Hypothesis Testing

Overview: Hypothesis Testing

- The concepts of ***randomness of a sample*** and the ***sampling distribution*** provide the ***link*** between a sample statistic and the underlying parameter of its parent population.
- The sample gives us a ***glance*** at the otherwise invisible and unknown parent population.
- ***Inferential statistics allows us to assess the likelihood*** whether a sample statistics deviates

  [a] more due to ***random chance*** from an ***assumed value*** of the underlying parent population parameter, or

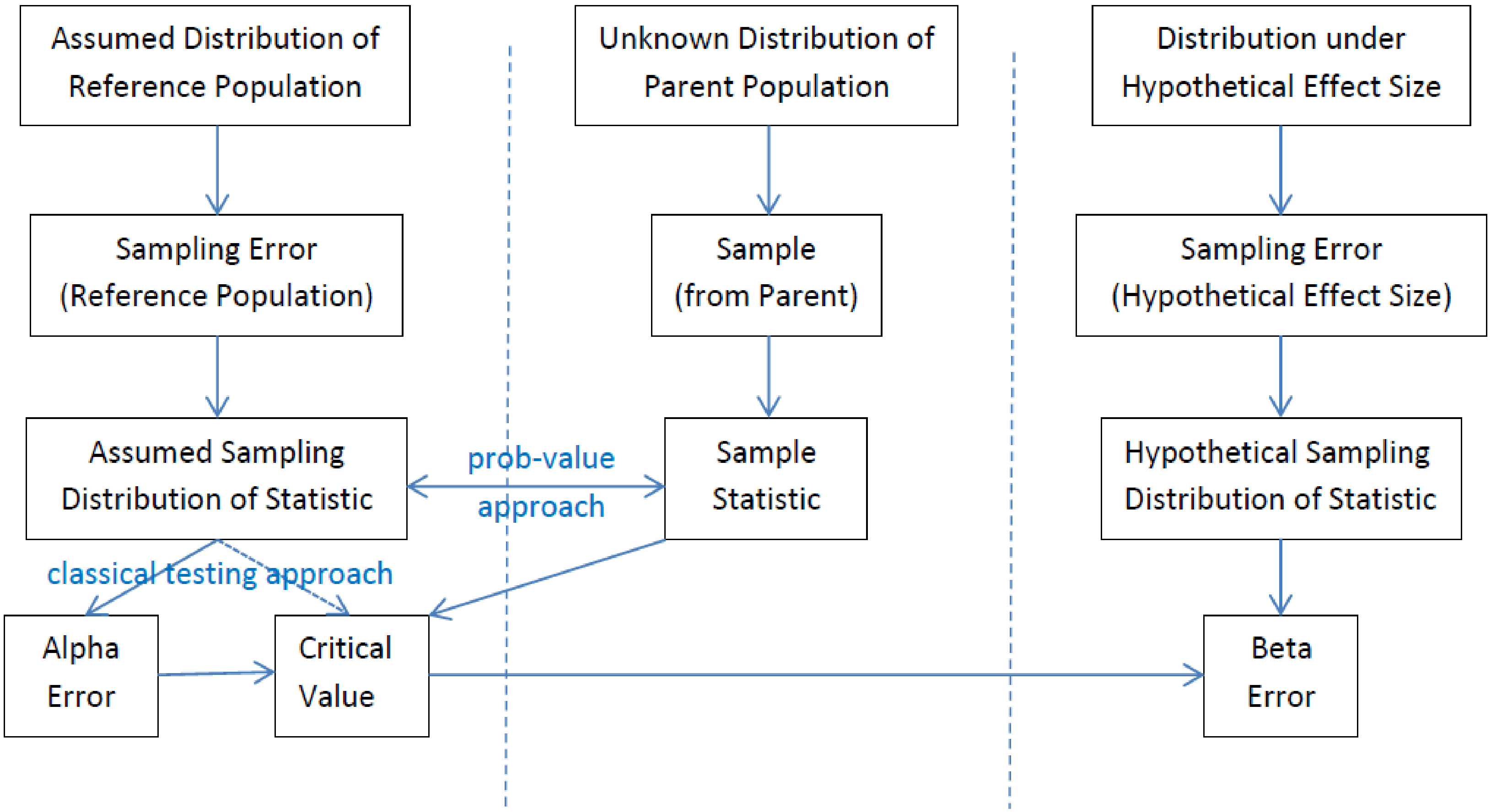
  [b] more due to ***some substantial differences*** from the assumed value of the underlying parent population parameter.

Conceptional Components of Statistical Hypothesis Testing

The flow chart below displays the general components of hypothesis testing. Each block starts with a particular, perhaps hypothetical, underlying population:

- We assume a hypothetical reference distribution for the unknown underlying parent population, i.e., given the ***null hypothesis would be true*** the reference distribution is ***equal*** to the parent distribution.
- We would like to make a statement about a parameter of the unknown parent population based on the drawn sample.

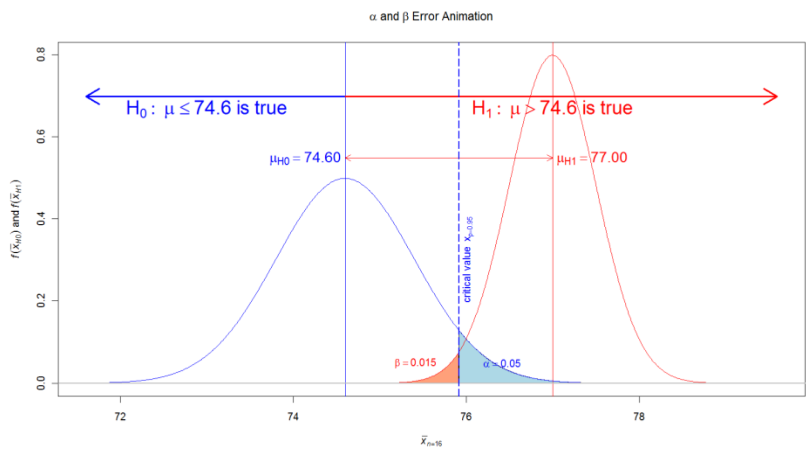
- A set of <u>hypothetical populations</u> ***assume a given set of possible population parameters*** and allow us to make ***what-if*** statements.



- Steps:
  1. We develop the sampling distribution of a test statistics assuming the sample was drawn from the <u>assumed reference distribution</u>.
  2. We draw a sample from the unknown <u>parent population</u> and calculate its associate sample statistic.
  3. We evaluate how likely this test statistic comes from the <u>assumed reference distribution</u>.

     $\Rightarrow$ Extreme sample statistics (in either tail of the reference sampling distribution) are unlikely to have been drawn from the reference distribution and lead with probability α to a rejection of the null hypothesis H_0.
- Example: The ***blue*** sampling distribution assumes the $H_0\colon \mu \leq 74.6$ is true. A test statistic beyond the critical value is highly unlikely to have been drawn for the reference population.

  If the true <u>parent population</u> has an expectation of $\mu_{H_1} = 77.0$ then the sampling distribution of the test statistic is shown in ***red***. (try the Shiny script **`app.R`** on your computer)



- Only the α-error can be controlled because it depends on the <u>assumed reference distribution</u>.
- However, the β-error cannot be controlled

because it depends on the unknown population distributions.

General Decision Errors:

- Possible outcomes of the decision making process about the true state of the parent population, given the observed sample is drawn from it, are:

Decision	**True State of the Parent Population**	
	H_0 is true (this is unknown to us because only the sample is available)	H_0 is false (this is unknown to us because only the sample is available)
Reject H_0 (based on sample)	Type I error: Controlled ***small sampling errors*** with probability $\boxed{\alpha}$ assuming the ***reference distribution is true***	Correct Decision: With probability $\boxed{1-\beta}$ depending on a ***given hypothetical state*** of the unknown population. This is also known as the ***Power*** *at a given effect size* $\left\|\mu_{H_0} - \mu_{H_1}\right\|$.
Fail to reject H_0 (based on sample)	Correct Decision: ***Tentatively*** $\boldsymbol{H_0}$ ***cannot be rejected***. However, if H_1 would be true, there is a possibility due to sampling variation not being able to reject H_0.	Type II error: Probability $\boxed{\beta}$ depending on a ***given hypothetical state*** μ_{H_1} of the unknown population the sample statistics falls into in ***non-rejection region*** due to sampling variation.

Statistical Inference: One Population Hypothesis Testing.

General steps of classical hypothesis testing

The steps are:

1. Formulation of the null and the alternative hypotheses and other assumptions.
2. Specification of the sample statistic and its statistical distribution under assumed reference distribution.
3. Selection of a level of significance (willingness to reject the null hypothesis even though it is correct due to having obtained an extreme sample).
4. Construction of a decision rule (i.e., critical values).
5. Collection of a sample and computation of the value of the test statistic.
6. Making a decision in favor or against the null hypothesis based on the value of the test statistic.

Formulation of the Hypotheses:

- Two-sided test for the point value of the hypothetical reference parameter θ:

  $H_0: \theta = \theta_0$ against $H_A: \theta \neq \theta_0$ or equivalently $H_A: \begin{cases} \text{either } \theta > \theta_0 \\ \text{or } \theta < \theta_0 \end{cases}$

  (two-sided because the true parameter can be left or right from θ_0)

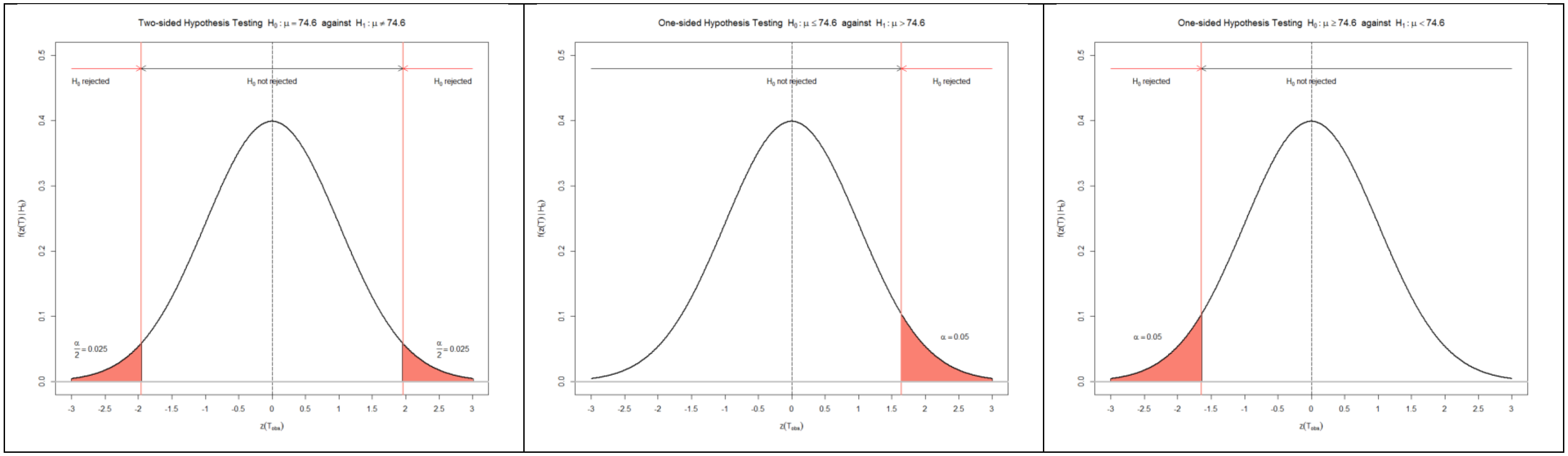
  This implies that we have two critical values, one on the left and one on the right of the sample distribution of the test statistic under the null hypothesis

- One sided for an interval of parameter values:

  $H_0: \theta \leq \theta_0$ against $H_A: \theta > \theta_0$ or

  $H_0: \theta \geq \theta_0$ against $H_A: \theta < \theta_0$

- It is advisable to review the diagram before selecting the direction of the hypotheses:



- Important notes:
    - In ***one-sided testing*** the null hypothesis is very ***specific*** by choosing a particular direction of deviation from the null hypothesis in favor of an alternative hypothesis.
    - The choice of the one- or two-sided specification depends on the problem under investigation. A ***one-sided hypothesis*** requires that the analyst is ***more knowledgeable*** about the underlying populations.
    - In a one-sided test the ***critical value moves closer*** towards the center of the test-statistics' distribution, because all error probability is concentrated in one tail: Therefore, if the direction of the alternative hypothesis has been ***guessed properly***, the null hypothesis will more likely be rejected.
    - Place the ***issue that you aim at confirming*** into the alternative hypothesis, because we only can control the α -error:
        - I.e., rejection of the null hypothesis leads to acceptance of alternative hypothesis.
        - The probability of erroneously rejecting the null hypothesis should be small.
        - However, the alternative hypothesis is usually less specific (broad range of possible parameters).
    - As a rule of thumb, ***choose the largest possible*** α ***-error*** that you can live with.
    - When H_0 is not rejected we need to say "***we fail to reject*** $\boldsymbol{H_0}$ ***at the error level*** $\boldsymbol{\alpha}$" rather than saying "*we accept* H_0 *at the error level* α" because there is always the possibility of committing a β error (i.e., H_0 is false but the observed sample is extreme

with regards to the unknown parent population, which leads to a failure of rejecting H_0).

- Remember: [a] assuming the null hypothesis is correct, then ***the variability of the test statistic*** only comes from ***sampling error***,
  [b] if the alternative hypothesis is correct, then the deviation of test statistic stems from a true ***difference to the expected value (the effect)*** under the null hypothesis ***plus a general sampling error***.

Example: Hypothesis Testing for the Population Proportion

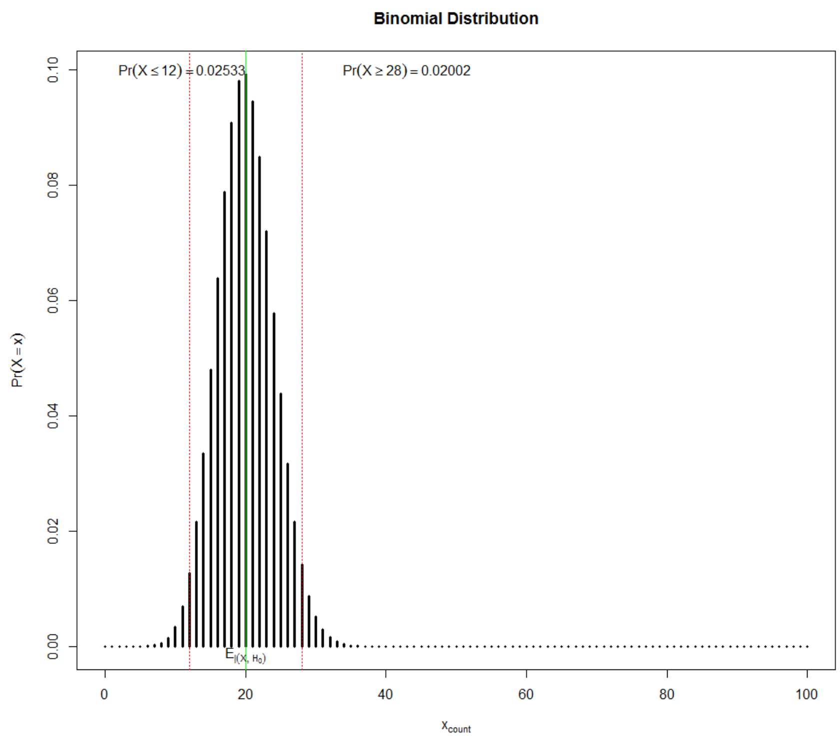
[a] Selection of a Sample Statistic and Its Sampling Distribution

- Based on the point estimator, e.g., $P = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ and its variances, we can design a ***test statistics***. Note that the variance must be evaluated ***assuming*** that null hypothesis correct, that is $H_0: \pi = \pi_0$.

  Thus the standard error for a binomial test scenario becomes $\sigma_P = \sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{n}}$

- We also can approximate the distributions under the null hypothesis of this ***test statistic*** for a sufficiently large sample size $n > 100$ by the normal distribution:
  $z(P) = \frac{P - \pi_0}{\sigma_P} \sim N(0,1)$.

- Excursion: For *smaller sample sizes* or extreme hypothetical parameters π_0 (leading to a skewed distribution of the test statistic) the exact binomial distribution should be used. However, one most likely ***will not be able to work with the exact Type I error probability α***, due to the discrete nature of the binomial distribution. See: **EXACTBINOMIALTEST.R**



- Notes:
  - Rather than showing the distribution of the rate estimator $\hat{\pi} = \frac{X}{n}$, the hypothetical distribution $X \sim Binomial(\pi_0 = 0.2, n = 100)$ of the ***equivalent count estimator*** $X = \sum_{i=1}^{n} X_i$ is used because it is binomial distributed.
  - The expected value under the null hypothesis is $E(X) = n \cdot \pi_0 = 20$
  - The critical values are $X_{lower} = 12$ and $X_{upper} = 28$ with their associated error probabilities $\alpha_{lower} = 0.02522$ and $\alpha_{lower} = 0.02002$, respectively.

    Note: Because the distribution is discrete the total error probability of $\alpha = 0.05$ cannot be exhausted.

[b] Selection of the Level of Significance

- Select the significance level α according to the amount of ***risk you are willing to take*** to reject the ***correct null hypothesis in error*** due to sampling variations.

[c] Construct the Decision Rule

- Split the parameter space of π in dependence of the significance level into an ***area in accordance with the null hypothesis*** and tail areas (two-sided tests) or one tail area (one-sided test) ***indicating the alternative hypothesis is more valid*** because of wrongly assumed null hypothesis. Here for a two-sided test:

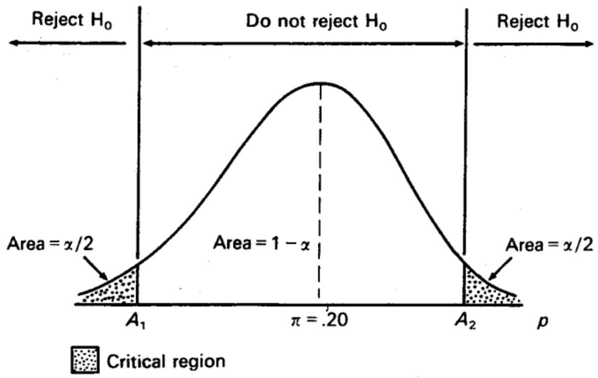


FIGURE 9-2. Sampling distribution of P, centered on hypothesized value $\pi = .2$.

- The selection of the acceptance and rejection areas depends on the specification of the null hypothesis as one- or two-sided.
- **Def. Critical region and critical values:** The critical region corresponds to those values of the test statistics, for which the null hypothesis is rejected.
  The limit(s) of the critical region are (is) the critical values.
- The two-sided null and the alternative hypotheses can be translated into
  - $H_0 : T \in \left]A_1, A_2\right[$ under the null hypothesis, and
  - $H_1 : T \leq A_1 \cup T \geq A_2$ under the alternative hypothesis: (these are the critical regions for a two-sided test)

[d] Example: residential mobility

- Assume that the observed value of for the sample proportion is $P = 0.26$

TABLE 9-4
Summary of Test of Hypotheses for Residential Mobility Example

Step 1. H_0: $\pi = .2$ and H_A: $\pi \neq .2$.

Step 2. P is chosen as the sample statistic.

Step 3. $\alpha = .05$.

Step 4. Reject H_0 if $p < .1216$ or $p > .2784$. (See Figure 9-3.)

Step 5. From random sample, $p = .26$.

Step 6. Because $.1216 < p < .2784$, do not reject H_0.

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.2(.8)}{100}} = .04$$

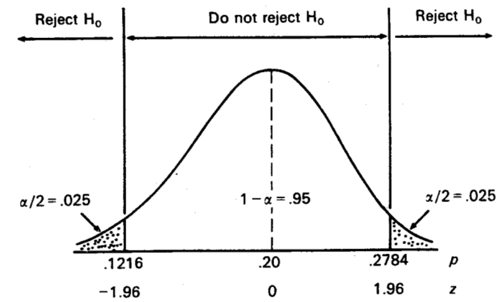


FIGURE 9-3. Determining critical region limits for residential mobility example, $\alpha = .05$.

Testing for population expectations

TABLE 9-5
Single-Sample Tests for μ

Background
A single random sample of size n is drawn from a population, and the sample mean $\bar{x}$ is calculated. The sample value will be compared to the hypothesized value to determine if H_0 should be rejected. The sampling distribution of $\bar{X}$ is known exactly if X is normal, which permits one to attach an exact PROB-VALUE to the sample result. If X is approximately normal, the PROB-VALUE is approximate.

Hypotheses

H_0: $\mu = \mu_0$	H_0: $\mu \geq \theta_0$	H_0: $\mu \leq \theta_0$
H_A: $\mu \neq \mu_0$	H_A: $\mu < \mu_0$	H_A: $\mu > \mu_0$
(*A*)	(*B*)	(*B*)
(two-tailed)	(one-tailed)	(one-tailed)

Test statistic
Case 1: σ known. The test statistic Z is normally distributed. Using the sample mean $\bar{x}$, the observed value is

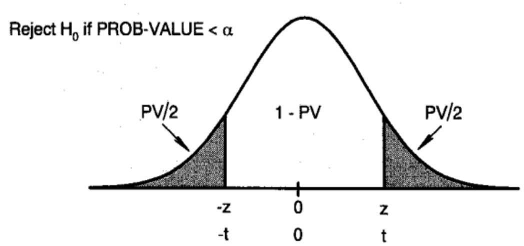
$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Case 2: σ unknown. The test statistic T is t-distributed with $n - 1$ degrees of freedom. Using the sample mean and standard deviation $(\bar{x}, s)$, the observed value is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

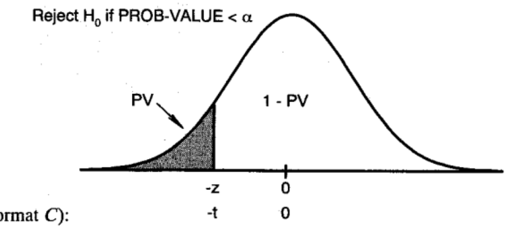
PROB-VALUE (PV) and Decision Rules
Two-tailed test (format *A*):

$$PV = P(|Z| > z) = P(Z < -z) + P(Z > z) \qquad PV = P(|T| > t) = P(T < -t) + P(T > t)$$



One-tailed test (format *B*):

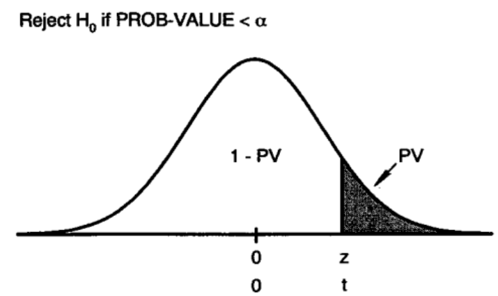
$$PV = P(Z < -z) \qquad PV = P(T < -t)$$



One-tailed test (format *C*):

$$PV = P(Z > z) \qquad PV = P(T > t)$$

(continued)



Relationship between Hypothesis Testing and Confidence Intervals

- The confidence interval at a given probability $1-\alpha$ complements the classical hypothesis testing approach for ***two-sided*** hypotheses testing.
- The confidence interval is constructed using the ***observed value of the test statisitic***.
- If the hypothetical value θ_0 under the null hypothesis falls ***within*** the confidence interval, then we are ***not able*** to reject the null hypothesis.

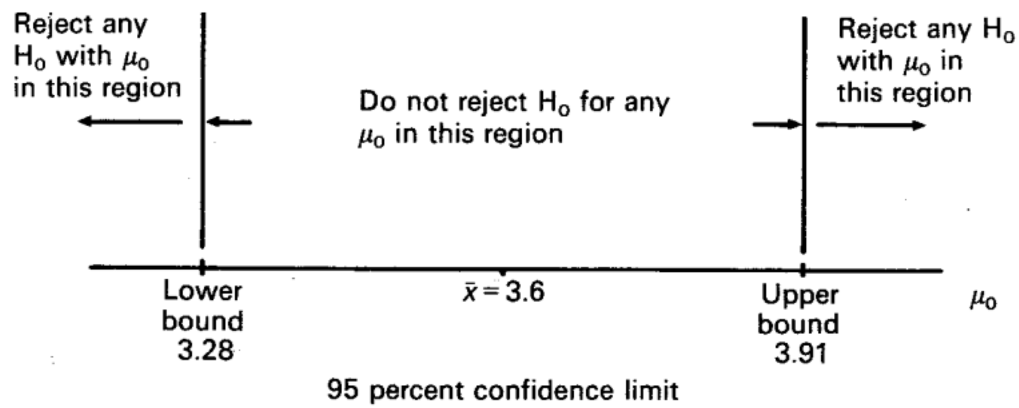


FIGURE 9-9. Relationship between confidence interval estimation and hypothesis testing for Example 9-1.

Problems of Significance Tests

- <u>Practical relevance:</u> A statistical significant result does not mean that it is relevant for practical purposes. Vice versa, a small but insignificant difference may be practically relevant.

Statistical significance is based on the distribution of the test statistic under the null hypothesis, whereas practical significance is based on the absolute difference of the test statistic from the value assumed under the null hypothesis.

This difference needs to be larger than a given threshold value to induce action.

Example: a home owner has a mortgage at 5.5% interest on his/her home. Another lender offers an interest rate of 5.3% to refinance the mortgage. Will the home owner take this offer because he/she saves substantial money after refinancing expenses?

- Deductive reasoning: A sample must always be drawn after we have formalized the hypotheses.

  This leads to problems for prior exploratory data analyses, which may bias the hypothesis formulation process

  The hypothesis would become dependent on a sample, if we use the same sample to formulate the hypothesis and therefore, predetermine the outcome.
- Large sample problem: As the sample size n becomes excessively large, virtually ***non-existing differences*** of the test statistic from the hypothetical parameter become statistically significant because the ***standard error*** of the test statistics shrinks toward zero.
- Multiple testing: The error probability increases if we perform multiple tests on the same dataset:

  Assume we perform two independent tests on the same data, the probability of not rejecting

[NR] both correct null hypotheses becomes

$$\Pr(NR \cap NR) = (1-\alpha)\cdot(1-\alpha) = (1-\alpha)^2 < (1-\alpha)$$

Therefore, the error probability of rejecting [R] at least on test incorrectly becomes

$$\Pr(R \cup R) = 1-(1-\alpha)^2 = \alpha_{new} \text{ with } \alpha_{new} > \alpha$$

Significance Test for the Correlation Coefficient

Test Under the Null Hypothesis of Independences (see BBR p486-487)

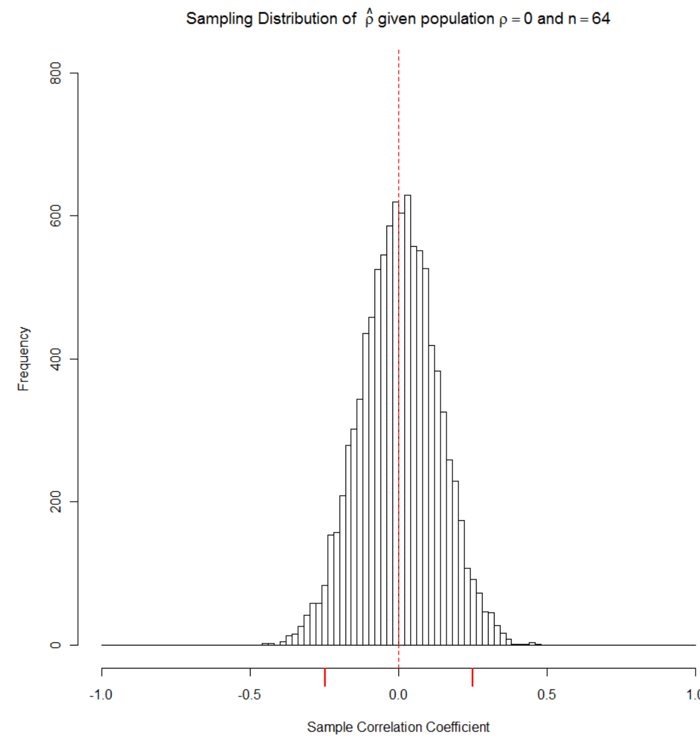
- Assumption: The ***bivariate reference population*** consists of two uncorrelated variables. $\Rightarrow$ This leads to the null hypothesis $H_0: \rho_0 = 0.0$ and the alternative hypothesis $H_1: \rho_0 \neq 0.0$. Assuming a correlation of zero under the null hypothesis is sensible because it is ***least specific*** and ***neutral*** with regards to a ***potential dependencies*** between two variables.
- The ***population correlation parameter*** is denoted by ρ whereas the ***sample statistic*** is denoted by r.
- If we can assume that both variables X_1 and X_2 are ***approximately jointly normal distributed and uncorrelated***, then we do not need to evaluate the sampling distribution with a simulation experiment. Its distribution is known:
  - The sample correlation coefficient r follows under the ***null hypothesis*** $H_0: \rho = 0$ a **t-*distibution*** with $df = n-2$ degrees of freedom where n is the number of

observations.

It has an expected value of $E(r) = 0$ and a standard error of $\sqrt{Var(r)} = \sqrt{\frac{1-r^2}{n-2}}$, that is

$$t = \frac{r - E(r)}{\sqrt{Var(r)}} \sim t_{df=n-2}$$

- We are ***losing two degrees of freedom*** for the correlation coefficient because for each variable first their means need to be estimated.



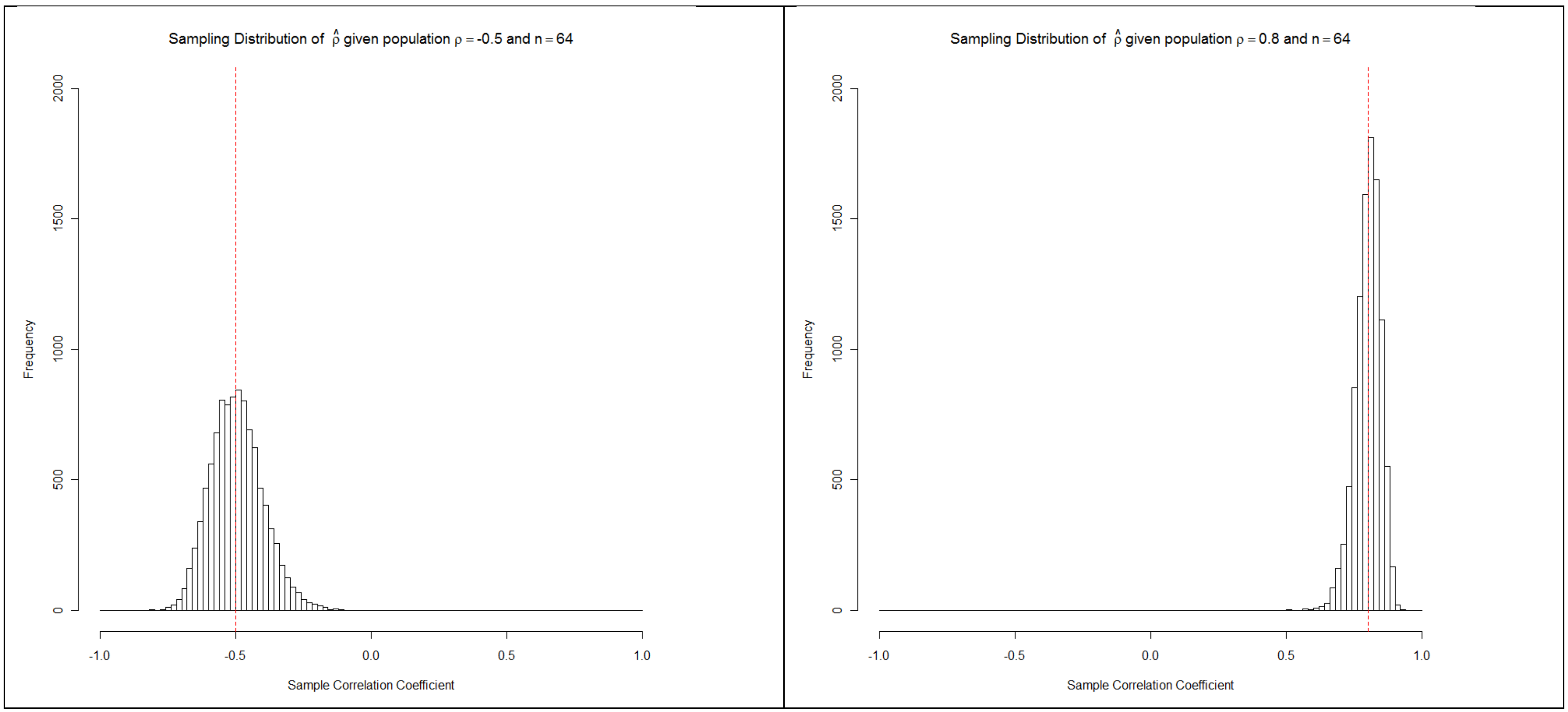
- The test statistic becomes

$$t = \frac{r - E(r)}{\sqrt{Var(r)}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}.$$

- Extreme values t indicate that the pairs of sample observations are most likely not originating from an uncorrelated bivariate parent population.
- For a sample sizes of more than $n > 30$ observation pairs the t-distribution can be approximated by the standard normal distribution.
- See two-sided test example with $\alpha = 0.05$ to the left:

Distribution of ρ under Alternative Hypotheses

- Under the assumption that $\rho_0 \neq 0.0$ both variables are correlated and the true population correlation coefficient will move
    - to either $\rho \rightarrow +1$ for positive correlation
    - and $\rho \rightarrow -1$ for negative correlation.
- Now the distribution can no longer be symmetric because it is bound asymmetrically:



- Clearly, neither a normal approximation nor the t-distribution can be used to evaluate the significance of an observed correlation coefficient of the critical values.
- However, the so-called Fisher z-transformation can be used to transform the correlation coefficient r to be approximately normal distributed:

$$z(r) = \frac{1}{2} \cdot ln\left(\frac{1+r}{1-r}\right) \text{ with}$$

$$E[z(r)] = \frac{1}{2} \cdot ln\left(\frac{1+\rho_0}{1-\rho_0}\right) \text{ and}$$

$$Var[z(r)] = \frac{1}{N-3}$$

Therefore,

$$\frac{z(r) - E[z(r)]}{\sqrt{Var[z(r)]}} \sim N(0,1)$$

is approximately standard normal distributed and can be used to test

$$H_0 : \rho = \rho_0 \text{ against } H_1 : \rho \neq \rho_0$$

- Example of the sample distribution after the Fisher z-transformation: