Hypothesis Testing

- We have hypotheses about the population, and we use a random sample to decide which is accurate.
- We may be testing a new medication/treatment, a new scientific theory, etc.
- Is the value of a parameter equal to a particular value, or is it larger? smaller? just different?
- In this course we will mainly consider the hypotheses about the population means and population proportions.

Null and Alternative Hypotheses

1. Null hypothesis:

   - Expressed in the form $\theta = \theta_0$, where θ is a population parameter and θ_0 is a specific value hypothesized for that parameter.
   - Denoted by H_0.
   - H_0 is the hypothesis to fall back on if we don't have enough evidence to support the other hypothesis (alternative hypothesis) H_a.

2. Alternative hypothesis:

   - Expressed in one of the three forms:

$$\left.\begin{array}{l}\theta > \theta_0 \ \ \text{(upper-tailed)} \\ \theta < \theta_0 \ \ \text{(lower-tailed)}\end{array}\right\} \text{one-tailed}$$
$$\theta \neq \theta_0 \ \ \text{(two-tailed)}$$

   - Denoted by H_a (or occasionally H_1).
   - Usually the objective is to show that H_a is true.

Types of Errors

Since we have two hypotheses, there are two errors that we could make.

	Actually true	
Decision	H_0	H_a
Reject H_0	Type I error	Correct decision
Not reject H_0	Correct decision	Type II error

1. Type I error:

   - Occurs if H_0 is rejected when it's really true.
   - Probability of making a type I error is denoted by α and called the *level of significance* (or *level*) of the test.

2. Type II error:

   - Occurs if H_0 is not rejected when it should be (when H_a is true).
   - Probability of making a type II error is denoted by β and $1-\beta$ is called the *power* of the test.

How to Minimize α and β?

- It isn't possible to reduce the probability of both types of error to 0 at the same time.
- For a fixed sample size, if we try to reduce α, then β goes up, and vice-versa.
- We fix the value of α and try to choose a decision rule such that β is minimized for that α. Often α is chosen to be 0.05.
- Statistical tests are conservative about H_0.

The Basic Ideas

- We try to reject H_0 by using the argument by contradiction.

- We try to see if observing the data, that we actually have, is improbable under the assumption of H_0. If it is, we reject H_0 in favor of H_a.

- If we do not get enough evidence against H_0, we do not reject H_0. That's why the statistical tests are conservative about H_0.

- For this reason, we usually say "we fail to reject H_0" instead of saying "we accept H_0."

Example: Medical Treatments

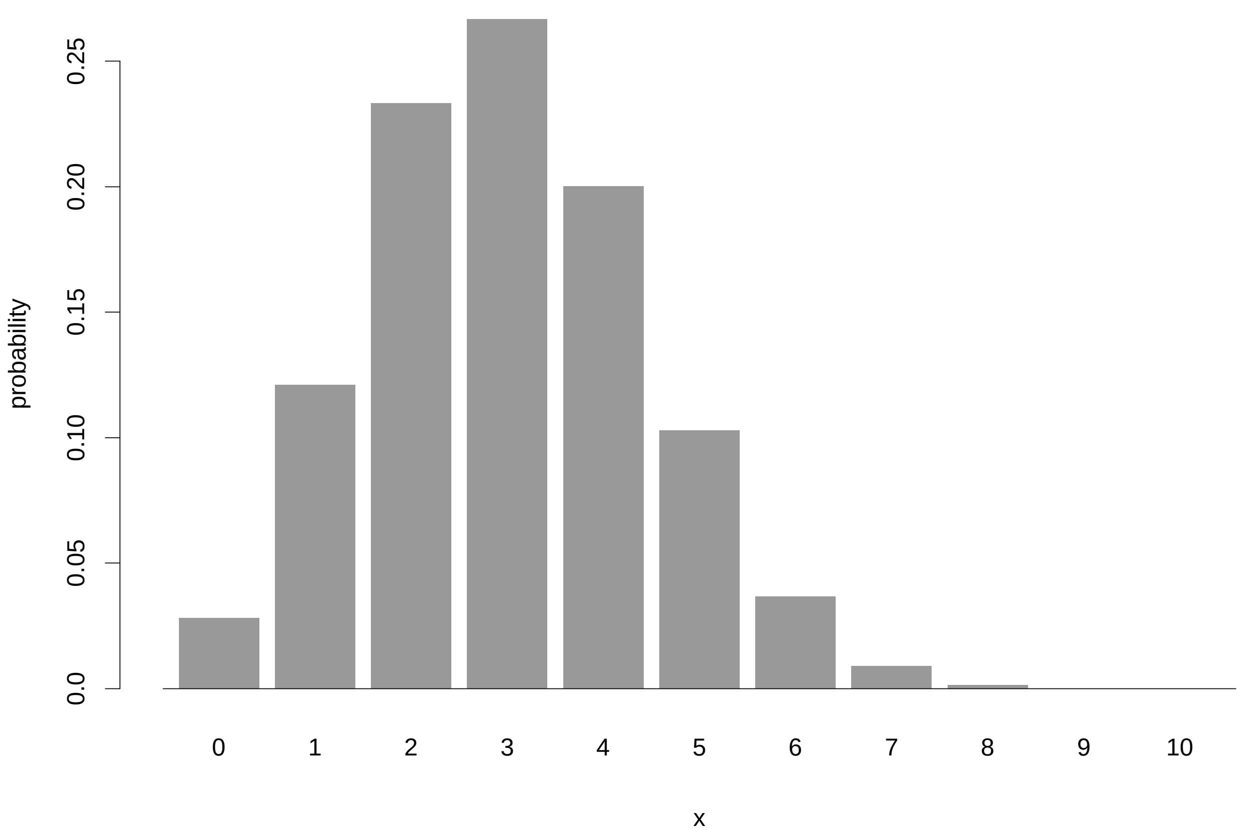
Experience has shown that 30% of people with a certain illness recover. We have developed a new medication, and we have given it to 10 randomly selected patients to test whether it increases the recovery rate.

1. What are the hypotheses that we want to test?

2. How many of them have to recover before we believe that the medicine increases the recovery rate? To answer this question, we want to look at the probabilities of 0, 1, 2, ... 10 patients recovering if the probability of recovery is unchanged by the new medicine.

3. If we assume that each person's probability of recovery is the same, the number of people recovering has binomial distribution with $n = 10$ and $p = 0.3$.

Binomial Probabilities for Our Example



Larger example

- Our previous example was a very small one, but the same ideas hold for more complicated problems.
- We usually have a much larger sample size, and we use the central limit theorem to approximate the sampling distributions with the normal distribution (or t distribution).

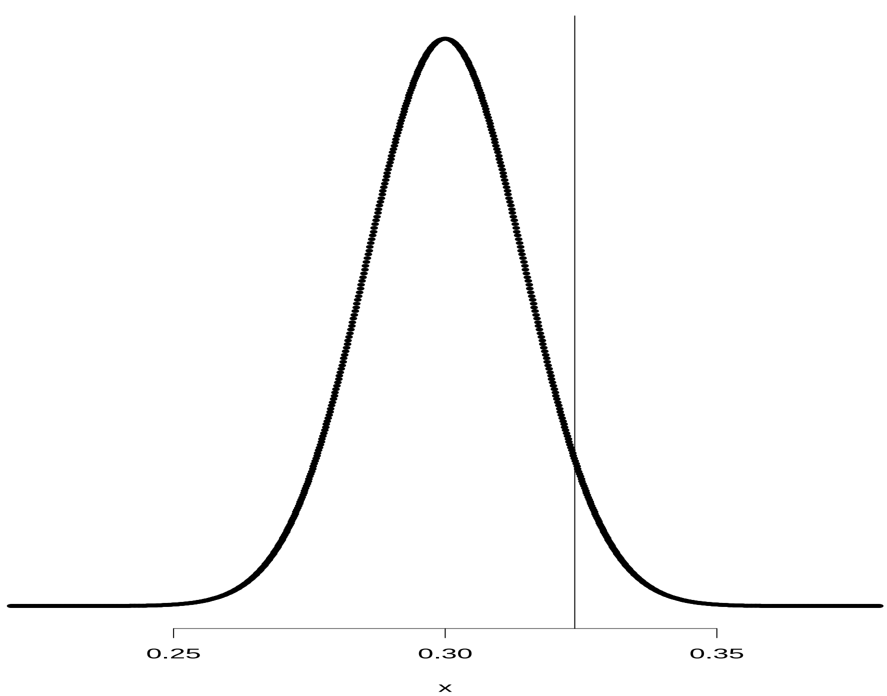


Figure 1: Approximate sampling distribution with n=1000

Test Statistic

- A test statistic is used to measure the difference between the data and what is expected under H_0.
- Test statistics are formed using estimators of the parameters about which the hypotheses are made.
- Since we will mainly consider the hypotheses about population means and proportions, we will consider the relevant test statistics based on the point estimators we have already seen.
- If the parameter of interest is θ, usually we use the test statistics of the form $\frac{\hat{\theta}-\theta_0}{SE_{\hat{\theta}}}$, where $\hat{\theta}$ is the unbiased point estimator for θ (what we have seen in chapter 8) and θ_0 is the value of θ under H_0. This form of test statistics will usually follow a normal or a t distribution.

Rejection Region

- A rejection region is a region such that if the test statistic falls in that region, we reject H_0.
- Usually α (the significance level of the test) is given. Then we have to find the rejection region such that the probability of observing the test statistic in that region, given that H_0 is true, is α.
- We need to consider the sampling distribution of the estimator under H_0, and determine the values that fall in the rejection region.

Large Sample α-level Tests

In the following discussion θ can be the population mean (μ), or population proportion (p), or difference between two population means ($\mu_1 - \mu_2$) or two population proportions ($p_1 - p_2$). For two sample tests, we assume independence between the samples.

Hypotheses: $H_0 : \theta = \theta_0$ vs $H_a : \theta \begin{smallmatrix} > \\ \neq \\ < \end{smallmatrix} \theta_0$.

Test statistic: $Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} \sim N(0, 1)$.

- $SE(\hat{\theta})$ is calculated under H_0 (especially important for testing a hypotheses about proportion).
- If population SD is unknown, use sample SD.

Rejection region:

$\{Z > z_\alpha\}$ for upper-tailed alternative

$\{Z < -z_\alpha\}$ for lower-tailed alternative

$\{|Z| > z_{\frac{\alpha}{2}}\}$ for two-tailed alternative

Small Sample α-level Tests

- We will further assume that the population is normal (for two-sample tests both populations are normal).
- In the following discussion θ can be the population mean (μ), or difference between two population means ($\mu_1 - \mu_2$).
- For two sample tests we assume equal population variances (if they are unknown) and independence between the samples.

There are two cases.

Case 1: When the population SD is known (or both population SD's are known for two sample tests), we can use exactly the same method as in the large sample case.

__**Case 2:**__ If the population SD is unknown, we use sample SD to estimate it, and we have the following test.

<u>Hypotheses:</u> $H_0 : \theta = \theta_0$ vs $H_a : \theta \substack{> \\ \neq \\ <} \theta_0$.

<u>Test statistic:</u>

1. When $\theta = \mu$, that is, one-sample case:

   $T = \frac{\bar{X} - \theta_0}{s/\sqrt{n}} \sim t$ with $n - 1$ d.f.

2. When $\theta = \mu_1 - \mu_2$, that is, two-sample case:

   $T = \frac{\bar{X}_1 - \bar{X}_2 - \theta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t$ with $n_1 + n_2 - 2$ d.f.

   where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

<u>Rejection region:</u>

$\{T > t_\alpha\}$ for upper-tailed alternative

$\{T < -t_\alpha\}$ for lower-tailed alternative

$\{|T| > t_{\frac{\alpha}{2}}\}$ for two-tailed alternative

The t distribution with appropriate d.f. is used to find t_α or $t_{\frac{\alpha}{2}}$.

Example: Medical Treatments (cont.)

Instead of randomly choosing 10 ill patients, we were able to randomly choose 1000. We observed that 350 of the 1000 recovered after being given the medicine. Is there evidence to lead us to believe that those given the new medicine have a higher rate of recovery than 30%? Test with significance level $\alpha = 0.05$.

Example: Bottling Factory

A bottling factory fills thousands of 20oz bottles daily with soda, but not all the bottles are filled to the same level. A random sample of bottles was taken from the factory line, containing the following amounts of soda (in oz):

19.8 20.1 19.7 19.2 19.9 20.0 19.8 19.9 19.7

Assuming that the distribution of amounts of soda is approximately normal, is there evidence to lead us to believe that the mean level of filling is not 20 oz? Test with significance level $\alpha = 0.05$.

From the sample data, we have: $n = 9$, $\bar{x} \approx 19.79$, $s \approx 0.26$

Attained Levels of Significance (p-values)

- If we assume that H_0 is true, we can calculate the probability of seeing a test statistic that is this far away from what we would expect given H_0, or further.

- This probability is the *p-value* or *attained level of significance.*

- If the p-value is less than α, this is equivalent to saying that the test statistic falls in the rejection region.

- Reporting the p-value lets each reader know whether the null hypothesis would be rejected using his/her own choice of α.

Example: Comparing Mean Incomes

We want to compare the mean family income in two states. For state 1, we had a random sample of $n_1 = 100$ families with a sample mean of $\bar{x}_1 = 35000$. For state 2, we had a random sample of $n_2 = 144$ families with a sample mean of $\bar{x}_2 = 36000$. Past studies have shown that for both states $\sigma = 4000$. Is the mean income in state 1 lower than that in state 2? Test at significance level $\alpha = 0.01$. Give the p-value.

Summary: Hypothesis Testing Methodology

1. Establish the null hypothesis.

2. Determine the alternative hypothesis $\rightarrow$ upper-tailed, lower-tailed or two-tailed test.

3. Define assumptions needed for the test, make sure they are met. (Ex: paired/unpaired, equal variances, etc.)

4. Calculate the test statistic:

$$Z/T = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$$

   where $\hat{\theta}$ is an unbiased point estimator for θ, θ_0 is the value of the parameter given by the null hypothesis and $SE(\hat{\theta})$ is calculated (under the null hypothesis, if information available) or estimated (from the data).

5. Based on α and H_a, determine the rejection region. This is the portion of the tail(s) (based on H_a) of the estimator's sampling distribution (under H_0) which has area α.

6. Determine the p-value (attained level of significance). This is the probability of seeing a result as extreme or more extreme than your test statistic under the null hypothesis, where "extreme" is defined by H_a.

7. Reject H_0 if p-value $< \alpha$ or the test statistic falls in the rejection region.

Example: Diff. in Means (Small Samples)

The strength of concrete depends, to some extent, on the method used for drying it. Two different drying methods yielded the results below for independently tested specimens (in psi). Do the methods appear to produce concrete with a statistically significant difference in mean strengths? Use $\alpha = 0.05$. Find the attained significance level.

Method 1 $n_1 = 7$ $\bar{y}_1 = 3250$ $s_1 = 210$

Method 2 $n_2 = 10$ $\bar{y}_2 = 3240$ $s_2 = 190$

Confidence Intervals vs. Hypothesis Tests

Imagine a two-tailed test with hypotheses $H_0 : \theta = \theta_0$ and $H_a : \theta \neq \theta_0$.

- For level of significance α, reject if test statistic (could be Z or T) falls in $Z < -z_{\frac{\alpha}{2}}$ or $Z > z_{\frac{\alpha}{2}}$.
- This means we fail to reject H_0 when $-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}$.
- We can re-write this:

  $-z_{\frac{\alpha}{2}} < \frac{\hat{\theta}-\theta_0}{SE_{\hat{\theta}}} < z_{\frac{\alpha}{2}}$

  $-\hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}} < -\theta_0 < -\hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$

  $\hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}} > \theta_0 > \hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$

  $\hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}} < \theta_0 < \hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$

- As long as θ_0 lies within bounds $(\hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}}, \hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}})$, we fail to reject.

- These are confidence interval boundaries for confidence coefficient $(1-\alpha)$!

- A confidence interval with confidence coefficient $(1-\alpha)$ can suffice for a 2-tailed test with significance level α. If interval doesn't include θ_0, then we can reject H_0.

Can show similar relationship for one-sided confidence intervals and one-tailed hypothesis tests.

Ex: Compare CIs and Hypothesis Tests

We have a random sample of checking account balances at each of two banks, and we're interested in testing whether the mean checking account balances are different at the two banks. Use $\alpha = 0.10$.

Bank 1 $n_1 = 12$ $\bar{x}_1 = 1000$ $s_1 = 150$

Bank 2 $n_2 = 10$ $\bar{x}_2 = 920$ $s_2 = 120$

More About Type II Error

- We've discussed the probability of type I error, α, associated with hypothesis tests, but we may also be interested in the probability of type II error, β.
- This is the probability of failing to reject H_0, when H_a is really true.
- Type II error is the probability that test statistic does not fall in rejection region, given that the parameter has a value categorized by H_a.
- In order to calculate this value, we need to calculate the test statistic given a specific value of the parameter under H_a.
- Power of the test = 1-β.

Example: Probability of Type II Error

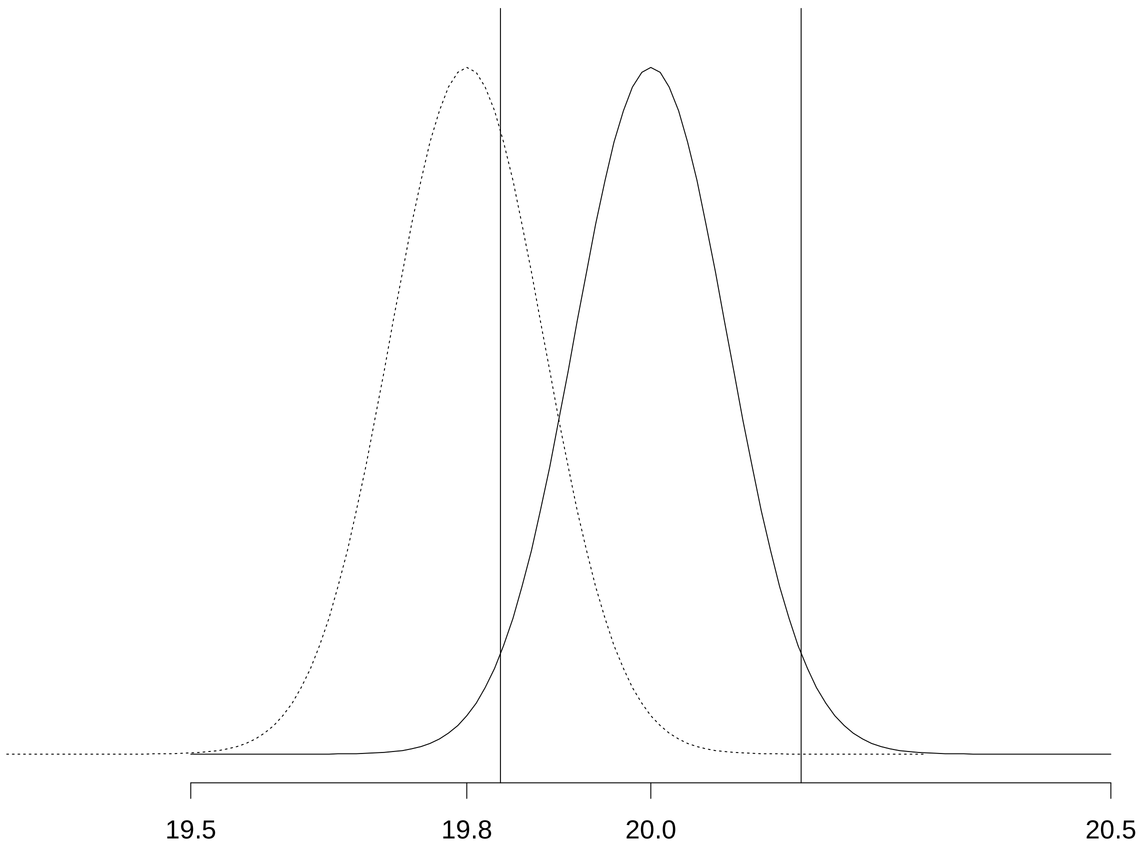
Experience has shown that 30% of people with a certain illness recover. Ten people are selected at random and given an experimental medicine; nine recover shortly thereafter.

Calculate β, the probability of type II error, with H_a: $p = 0.6$ and rejection region $\{Y \geq 8\}$.

Example: Probability of Type II Error

A bottling factory fills thousands of 20oz bottles daily with soda, but not all the bottles are filled to the same level. A random sample of $n = 9$ bottles was taken from the factory line to try to determine whether the mean filling level is 20oz or not. If we test these hypotheses with $\alpha = 0.05$, find the probability of type II error if $H_a : \mu = 19.8$. Assume that the distribution of amount of soda is approximately normal and that $\sigma = 0.25$ (known from previous experience).

Figure Illustrating Regions in Bottling Example



Example: Probability of type II error

A manufacturer claims that more than 20% of the public prefers her product. A sample of 500 people is taken to check her claim. If we test her claim with $\alpha = 0.05$, what is β if the percentage of the public preferring her product is really 22%?