**MATH 321: Mathematical Statistics**

# Lecture 7: Hypothesis Tests

Chapter 8: Tests of Statistical Hypotheses (8.1 - 8.3)

## Introduction

- Recall that the objective of statistics often is to make inferences about unknown population parameters based on information contained in sample data.

  These inferences are phrased in one of two ways:

  - As estimates of the respective parameters (point estimation / confidence intervals)
  - Or as tests of hypotheses about their values

- Hypothesis tests are essentially the scientific method viewed through statistics.
  - The scientist poses a hypothesis concerning one or more population parameters (e.g. that they equal specified values).
  - Then samples the population and compares observations with the hypothesis.
  - If the observations disagree with the hypothesis, the scientist rejects it.

    If not, the scientist concludes either that the hypothesis is true or that the sample did not detect the difference between the real and hypothesized values of the population parameters.

- Hypothesis tests are done in almost all fields where we are testing theory against observation. Examples:
  - A medical researcher may hypothesize that a new drug is more effective than another in combating a disease.

    To test her hypothesis, she randomly selects patients infected with the disease and randomly divides them into two groups: Group A gets the current drug and Group B gets the new drug.

    Then, based on the number of patients in each group who recover from the disease, the researcher must decide whether the new drug is more effective than the old.
  - A quality control engineer may hypothesize that a new assembly method produces only 5% defective items.
  - An educator may claim that two methods of teaching reading are equally effective.

- Statistics and what we will learn is what measures to take on the sample, how do make the decision of accept vs reject, what are the probabilities we made the correct / incorrect decision, etc.

## Elements of a statistical test

Hypothesis test overview

- Definition: A **hypothesis testing procedure or hypothesis test** is a rule that specifies
  - For which sample value the decision is made to reject $H_0$ in favor of $H_A$.
  - For which sample value the decision is made to "not reject" $H_0$ in favor of $H_A$.

- Any statistical test of hypotheses works in exactly the same way and is composed of the same essential elements.
  1. Null hypothesis $H_0$ and Alternative hypothesis $H_A$
  2. Test Statistic TS and Rejection Region RR
  3. Conclusion

- Example setup: Let $X$ equal the breaking strength of a steel bar. A company uses process I to manufacture steel bars and it is known that under process I, $X \sim \text{Normal}\,(\mu = 50, \sigma^2 = 36)$.

  The company wishes to test a new process, process II, and it is hoped that under process II $X \sim \text{Normal}\,(\mu = 55, \sigma^2 = 36)$.

- Hypotheses
  - Definition: A **hypothesis** is a statement about a population parameter.
  - The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.
    - The **Null hypothesis $H_0$** is an assumption about $\theta$ that is assumed to be _____.
    - The **Alternative hypothesis $H_A$** (or $H_1$, also called research hypothesis) is the _____ of the null hypothesis. The goal is generally to obtain evidence in favor of this.
  - Continuing steel bar example:
  - These are called **simple hypotheses** because each completely specifies the distribution of $X$. Could test $H_0$ against a **composite hypotheses**, which contains many possible alternative distributions.
  - In general, we have the following hypotheses:

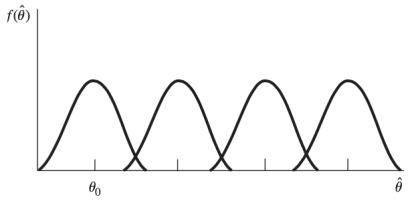
- Examples: (1) Define the parameter of interest and (2) state the null and alternative hypotheses and the directionality of the test (two-tailed, left-tailed or right-tailed) for the following scenarios:

  (a) A company reports that last year 40% of their reports in accounting were on time. From a random sample this year, they want to know if that proportion has changed.

  (b) Last year, 42% of the employees enrolled in at least one wellness class at the company's site. Using a survey from randomly selected employees, they want to know if a greater percentage is planning to take a wellness class this year.

  (c) There are two political candidates, and one wants to know from the recent polls if she is going to win a majority of votes in next week's election.

- Test statistic and rejection region
  - These are all about distributions of estimators based on assumptions from the hypotheses.

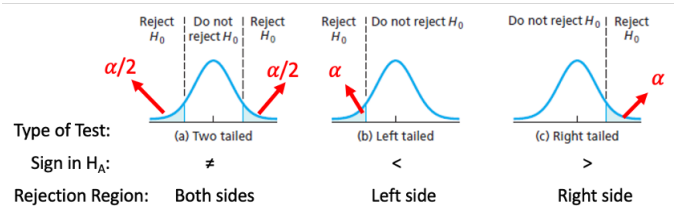



  - For example, for a right-tailed test
    - If $\hat{\theta}$ is close to $\theta_0$, it seems reasonable to accept $H_0$.
    - If in reality $\theta > \theta_0$, then $\hat{\theta}$ is more likely to be large.

      Consequently, large values of $\hat{\theta}$ (relative to $\theta_0$) favor rejection of $H_0 : \theta = \theta_0$ and acceptance of $H_A : \theta > \theta_0$.

- Simply stated, we have to determine when there is or is not enough _____ against the _____ based on our _____.

  In other words, which tail do we make the conclusion of reject, which comes from the direction in the $H_A$, and how large is the area.




- The hypothesis test is specified in terms of the test statistic and the corresponding rejection region.
  - **Test statistic (TS)** is a function of the sample $W(X_1, \ldots, X_n)$, think of this as the point estimator $\hat{\theta}$.
  - **Rejection Region (RR)** (or critical region) is the subset of the sample space (range of sample) for which $H_0$ will be rejected. RR is defined with the TS (these two parts are always together).
- Once these are defined, hypothesis tests are really easy; we then just observe data and see where it falls.
- In general, we can state the rejection region as

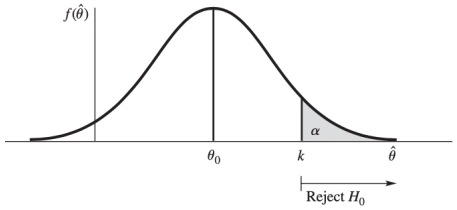
  $RR = \{\text{set of } (x_1, \ldots, x_n) \text{ such that (some math statement about TS } W(X_1, \ldots, X_n)\}$

- Continuing steel bar example: Suppose $n = 16$ bars were tested, intuitively we could choose a RR where larger values lead to rejecting $H_0$, say $RR = \{\bar{x} : \bar{x} \geq 53\}$.

- But how did we choose the value of $k$? More generally, how can we find some objective criteria for deciding which value of $k$ specifies a good rejection region of the form $\{\bar{x} \geq k\}$?

- Significance level
  - The **significance level** $\alpha$ of the test is what determines how large the RR is and represents the probability of rejecting the null hypothesis.

    The actual value of $k$ is chosen by fixing this and finding $k$ accordingly.
  - Recall under the null hypothesis, the distribution of $\hat{\theta}$ is known. So we can find $k$ such that (for example with a right-tailed test):




- The significance level is chosen before running the test. Setups will say something similar to: "Determine if there is enough evidence at the 5% significance level."

**Building hypothesis tests**

Hypothesis test setup

- Just like with confidence intervals, all of the hypothesis tests we will build start from this general setup and use properties of normal distributions or the central limit theorem to get the test statistic and rejection region of interest.

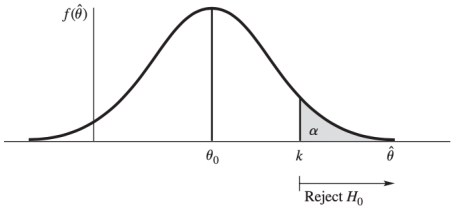
- For hypothesis tests, we will consider same variables that affect the formation of our confidence intervals:
  - Independent or dependent samples
  - Sample sizes $n_1$ and $n_2$ (large or small)
  - Population distributions $X_1$ and $X_2$ (normal or not normal)
  - Population variances $\sigma_1^2$ and $\sigma_2^2$ (known or unknown and ratio of variances)

Large sample tests

- Setup: Suppose we want to test a set of hypotheses concerning a parameter $\theta$ based on a random sample(s) $X_1, \ldots, X_n$. Additionally, let the estimator $\hat{\theta}$ have an (approximately) normal sampling distribution with mean $\theta$ and standard error $\sigma_{\hat{\theta}}$.

- Then we have the following:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta > \theta_0$$
$$TS :$$
$$RR :$$




- Defining the RR (i.e. finding $k$)

  Assuming $H_0$ is true, if we desire an $\alpha$-level test, then

- Thus, an equivalent form of the test, with level $\alpha$ is:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta > \theta_0$$
$$TS :$$
$$RR :$$

- We can use this generalization for all of the tests that large sample tests we will do, and we can state the test statistic as

$$Z = \text{————}$$

  and thus they all have equivalent form of the rejection region (because the TS has been standardized).

- Conclusions and interpretations
  - Conclusions and interpretations (two steps) for hypothesis tests can follow a general format:

    Because our test statistic (COMPARISON of TS and RR) (IS or IS NOT) in the rejection region we (REJECT or FAIL TO REJECT) the null hypothesis.
    At the (ALPHA) significance level, there (IS or IS NOT) sufficient evidence to conclude (THE ALTERNATIVE HYPOTHESIS).

- Examples

  1. A honey farmer collects 55 ml of honey on average from each of his hives during summer months. Further, he knows that the amount collected from each hive is normally distributed with a variance of $\sigma^2 = 100$. This summer he is feeding his bees a new type of pollen and he suspects that it is causing them to produce more honey. A random sample of $n = 52$ hives yields $\bar{x} = 57.25$. Test the farmer's hypothesis at a significance level of $\alpha = 0.05$.

  2. A vice president in charge of sales for a large corporation claims that salespeople are averaging no more than 15 sales contacts per week. (He would like to increase this figure.) As a check on his claim, $n = 36$ salespeople are selected at random, and the number of contacts made by each is recorded for a single randomly selected week. The mean and variance of the 36 measurements were 17 and 9, respectively. Does the evidence contradict the vice president's claim? Use a test with level $\alpha = 0.025$.

3. A machine in a factory must be repaired if it produces more than 10% defectives among the large lot of items that it produces in a day. A random sample of 100 items from the day's production contains 15 defectives, and the supervisor says that the machine must be repaired. Does the sample evidence support his decision? Use a test with level $\alpha = 0.01$.

- Here is a summary of the large-sample $\alpha$-level hypothesis tests:

**Large-Sample $\alpha$-Level Hypothesis Tests**

$H_0 : \theta = \theta_0$.

$$H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative).} \\ \theta < \theta_0 & \text{(lower-tail alternative).} \\ \theta \neq \theta_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$.

$$\text{Rejection region: } \begin{cases} \{z > z_\alpha\} & \text{(upper-tail RR).} \\ \{z < -z_\alpha\} & \text{(lower-tail RR).} \\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR).} \end{cases}$$

- In any particular test, only one of the listed alternatives $H_A$ is appropriate. Whatever alternative hypothesis that we choose, we must be sure to use the corresponding rejection region.

  The correct one depends on the research question / goal: what are we trying to show or find evidence for?

- More examples

4. A psychological study was conducted to compare the reaction times of men and women to a stimulus. Independent random samples of 50 men and 50 women were employed in the experiment. The results are shown below. Do the data present sufficient evidence to suggest a difference between true mean reaction times for men and women? Use $\alpha = 0.10$.

| Men | Women |
|---|---|
| $n_1 = 50$ | $n_2 = 50$ |
| $\bar{y}_1 = 3.6$ seconds | $\bar{y}_2 = 3.8$ seconds |
| $s_1^2 = .18$ | $s_2^2 = .14$ |

5. A car manufacturer aims to improve the quality of the products by reducing the defects and also increase the customer satisfaction. Therefore, he monitors the efficiency of two assembly lines in the shop floor. In line A there are 18 defects reported out of 200 samples. While the line B shows 25 defects out of 600 cars. At $\alpha = 5\%$, are the differences between two assembly procedures significant?

Small sample tests for $\mu$ and $\mu_1 - \mu_2$

- If we are testing one or two population means and the sample size is not large enough so that $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}} \overset{approx}{\sim} \text{Normal}\,(0, 1)$, then we need a different procedure.
- Just like with confidence intervals, we can switch to procedures based on the $t$-distribution when sampling from Normal distribution(s) (assuming unknown equal variances of both populations).

  The process is the same as the large sample $Z$-tests shown previously. We are just standardizing the point estimator and rearranging to get the rejection region, except it is based on $t$ critical values now.
- If $H_0 : \mu = \mu_0$ is tested against $H_A : \mu < \mu_0$ then

- Here is a summary of the small-sample $\alpha$-level tests for $\mu$

**A Small-Sample Test for $\mu$**

Assumptions: $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from a normal distribution with $E(Y_i) = \mu$.

$H_0 : \mu = \mu_0$.

$$H_a: \begin{cases} \mu > \mu_0 & \text{(upper-tail alternative).} \\ \mu < \mu_0 & \text{(lower-tail alternative).} \\ \mu \neq \mu_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $T = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}}$.

$$\text{Rejection region:} \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases} \quad t_\alpha, \text{ with df} = n - 1$$

- If we are testing two independent means $\mu_1 - \mu_2$ and assume both Normal distributions with common unknown variance $\sigma^2$, then we use the pooled variance $S_p^2$ as the estimator for $\sigma^2$ in the standard error $\sigma_{\bar{X}_1 - \bar{X}_2}$. Then

**Small-Sample Tests for Comparing Two Population Means**

Assumptions: Independent samples from normal distributions with $\sigma_1^2 = \sigma_2^2$.

$H_0 : \mu_1 - \mu_2 = D_0$.

$$H_a : \begin{cases} \mu_1 - \mu_2 > D_0 & \text{(upper-tail alternative).} \\ \mu_1 - \mu_2 < D_0 & \text{(lower-tail alternative).} \\ \mu_1 - \mu_2 \neq D_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $T = \dfrac{\overline{Y}_1 - \overline{Y}_2 - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$.

$$\text{Rejection region: } \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases}$$

Here, $P(T > t_\alpha) = \alpha$ and degrees of freedom $\nu = n_1 + n_2 - 2$.

- If we are testing dependent means $\mu_1 - \mu_2$, then

  This is called a **paired *t*-test**.

- Examples

  1. $X$ is the growth in an induced tumor in type of lab mice. It is known that the mean growth of the tumor without treatment is 4.0 mm and that the distribution of $X \sim \mathrm{N}(\mu, \sigma^2)$. Scientists believe a new type of enzyme will have an effect on the growth of the tumor. Scientists apply the enzyme to a random sample of $n = 9$ lab mice with the induced tumor and observe $\bar{x} = 4.2824$ and $s = 1.2$.

     Test the scientists' hypothesis at a significance level of $\alpha = 0.10$.

2. Data on the length of time required to complete an assembly procedure using each of two different training methods is shown below. Is there sufficient evidence to indicate a difference in true mean assembly times for those trained using the two methods? Test at the $\alpha = .05$ level of significance.

| Standard Procedure | New Procedure |
|---|---|
| $n_1 = 9$ | $n_2 = 9$ |
| $\bar{y}_1 = 35.22$ seconds | $\bar{y}_2 = 31.56$ seconds |
| $s_1^2 = 24.445$ | $s_2^2 = 20.0275$ |

P-values

- So far, we have only made the decision to reject or fail to reject based on whether or not the test statistic falls in the rejection region ($TS \stackrel{?}{\in} RR$). This is called the **traditional method**.
- Lets review some examples:
  - Example 1 (average honey): $TS : z = 1.622$ and $RR : \{Z > 1.645\}$ for $\alpha = 0.05$.
  - Example 3 (proportion of defectives): $TS : z = 1.667$ and $RR : \{Z > 2.362\}$ for $\alpha = 0.01$.
- In both of these, we made the conclusion to _____, but were "closer" to rejecting $H_0$ example ___. We can think of this as being a "stronger" result (i.e. more evidence against $H_0$ (just not enough), but we need a way to quantify the "strength" of the result independent of the significance level.

- Definition: A **p-value** is the probability that under the null hypothesis the test statistic will be at least as "extreme" as the observed value.

- Notes:
  - At least as "extreme" just means in the direction of the alternative hypothesis.

$$H_A: \qquad \theta < \theta_0 \qquad \theta > \theta_0 \qquad \theta \neq \theta_0$$

– Interpretation of p-values: The smaller the p-value becomes, the more compelling is the evidence that the null hypothesis should be rejected.

  For small p-values, think: If $\theta_0$ was true, the result we got had such a tiny probability to occur. So the original assumption of $\theta_0$ must not actually be true.

– Making the decision to reject or fail to reject based on whether or not the p-value is less than the significance level is called the **p-value method**.

– More formally, a p-value represents the smallest level of significance $\alpha$ for which the observed data indicate that the null hypothesis should be rejected; p-value is the **attained (observed) significance level**.

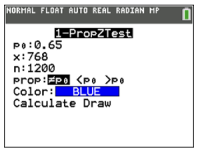
  Treating it like this leaves it up to the reader to evaluate the extent to which the observed data disagree with the null hypothesis and make their own choice in $\alpha$ in deciding whether or not to reject $H_0$. This is one advantage of p-values, and is why most scientific journals require p-values for all of their studies.

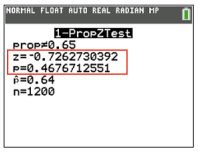
  (often $\alpha = 0.1, 0.05, 0.01$ are chosen out of convenience rather than a well-thought out choice.)

Calculator session

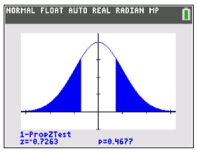
STAT > TESTS >

**One proportion test: $p$**




Calculate Output
prop = Alternative hypothesis
z = TS
p = p-value
$\hat{p}$= sample proportion
n = sample size

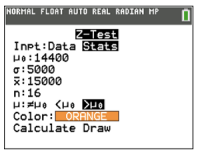


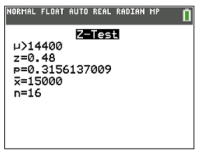

Draw Output
Plot (and displays values) of p = p-value and z = TS on the standard normal curve

$H_0$: $p = 0.65$
$H_A$: $p \neq 0.65$

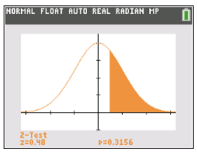
*This does NOT give us the Critical Value $z_{\alpha/2}$, we have to figure that out ourselves

**One mean test: $\mu$**



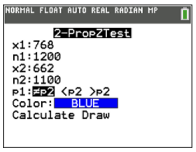

Calculate Output
$\mu$= Alternative hypothesis
z = TS
p = p-value
$\bar{x}$= sample mean
n = sample size

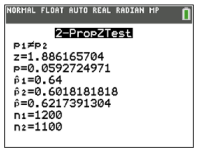



$H_0$: $\mu = 14{,}400$
$H_A$: $\mu > 14{,}400$

STAT > TESTS >

Two proportions test: $p_1 - p_2$




$H_0: p_1 - p_2 = 0$

$H_A: p_1 - p_2 \neq 0$

Calculate Output

$p_1 \neq <> p_2$ Alternative hypothesis

z = TS

p = p-value

$\hat{p}_1$ = sample proportion 1

$\hat{p}_2$ = sample proportion 2

$\hat{p}$ = pooled sample proportion → $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$

$n_1$ = sample size 1

$n_2$ = sample size 2

Relationship between confidence intervals and hypothesis tests

- For every confidence interval, there is an equivalent hypothesis test (and vice versa); two different ways of looking at the same thing.
- Demo for two-sided CIs and two-tailed tests:

  $100(1-\alpha)\%$ CI for $\theta$ $\qquad$ $\alpha$-level test $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$

- The complement of the rejection region is the **acceptance region AR**:

- Thus, we "accept" $H_0 : \theta = \theta_0$ if $\theta_0$ falls _____ the $100(1-\alpha)\%$ CI and reject if _____.

  So the confidence interval can be thought of as the set of values of $\theta_0$ for which $H_0 : \theta = \theta_0$ is "acceptable" at level $\alpha$.

  Based on this perspective, we can see that it is a range, not *one specific acceptable value* for the parameter. This is why we prefer to say "fail to reject" rather than "accept" the null hypothesis.

- One sided intervals and tests

  For $H_0 : \theta = \theta_0$ with level $\alpha$ and $100(1-\alpha)\%$ CIs

  - Upper tail test: $H_A : \theta > \theta_0$ $\qquad$ Reject if outside

  - Lower tail test: $H_A : \theta < \theta_0$

Errors in hypothesis tests

- In deciding to reject or fail to reject $H_0$, an experimenter might be making a mistake (we can never really know what the truth is, just like with confidence intervals).

  Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes.

- For any fixed rejection region, two types of errors can be made in reaching a decision.
  - **Type I error** is made if $H_0$ is rejected when $H_0$ is true.

    The probability of a type I error is denoted by $\alpha$, the **significance level** of the test.

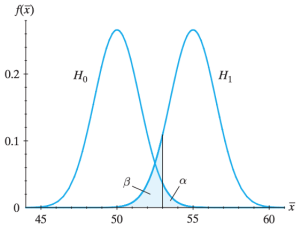
  - **Type Il error** is made if $H_0$ is accepted when $H_A$ is true.

    The probability of a type II error is denoted by $\beta$.

- We can think of $\alpha$ and $\beta$ as measuring the risks associated with the two possible incorrect decisions that might result from a statistical test. Because of this, they provide a very practical way to measure the goodness of a test.
- In calculating these error probabilities, we want tests that minimize both quantities while at the same time maximizing the **power** $= 1 - \beta$.

  The power of a test represents the probability of correctly rejecting a false null hypothesis (given a particular alternative hypothesis).

- Example: Find the probabilities of a type I error, type II error, and power for the breaking strength example (testing $H_0 : \mu = 50$ vs $H_A : \mu = 55; n = 16, \sigma^2 = 36$).



pdf of $\overline{X}$ under $H_0$ and $H_1$

- Note that the value of $\beta$ depends on the true value of the parameter $\theta$ in the alternative hypothesis (needed to assume $\mu = 55$ in the type II probability calculation).

  The larger the difference is between $\theta$ and the (null) hypothesized value of $\theta = \theta_0$, the smaller is the likelihood that we will fail to reject the null hypothesis.

  Example: Find the new type II error probability and power if $H_A : \mu = 57$ and if $H_A : \mu = 51$.

- This example shows that the test using $RR = \{\bar{x} \geq 53\}$ guarantees a low risk of making a type I error _____, but it does not offer adequate protection against a type II error (high $\beta$s with some alternative hypotheses).
- Typically, in practice the type I error probability (significance level) is controlled, and then we choose a test that minimizes the type II error probability (and thus maximizing the power).

  However, there is often some give and take with these: as one error likelihood decreases, the other often increases (i.e. $\alpha$ and $\beta$ are inversely related).
- So how can we improve our test? One way is to balance $\alpha$ and $\beta$ by changing the rejection region, specifically we can enlarge the RR.

  This will lead us to reject $H_0$ more often, which means accept $H_0$ less often.

- Often we have to think about the consequences of committing each type of error and determine which error is more severe and therefore how to minimize its probability.
- Example: Write the consequences of each type of error and determine which is more severe.

  All commercial elevators must pass yearly inspections. An inspector has to choose between certifying an elevator as safe (no repairs needed) or saying that the elevator is not safe (repairs are needed). There are two hypotheses:

  $$H_0 : \text{The elevator is not safe (repairs are needed)}$$
  $$H_A : \text{The elevator is safe (no repairs needed)}$$

  – Consequences of Type I error:

  – Consequences of Type II error:

- How can we reduce both? For almost all statistical tests, if $\alpha$ is fixed at some acceptably small value, $\beta$ decreases as the sample size increases.

  Intuitively obvious, collect more data!