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 θ that is assumed to be true. (status quo)
    * The **Alternative hypothesis H_A** (or H_1, also called research hypothesis) is the complement of the null hypothesis. The goal is generally to obtain evidence in favor of this.
  - Continuing steel bar example: $H_0: \mu = 50$ vs $H_A: \mu = 55$

    $X \sim N(50, 36)$ $\quad$ $X \sim N(55, 36)$

  - 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. Ex $H_A: \mu > 50$
  - In general, we have the following hypotheses:

H_0	H_A
$\theta = \theta_0$	$\theta \neq \theta_0$, $\theta = \theta_1$
$\theta = \theta_0$	$\theta > \theta_0$
$\theta = \theta_0$	$\theta < \theta_0$

- 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.

  → Let p= true proportion of on-time reports

  → H_0: $p = 0.4$

  H_A: $p \neq 0.4$ ⇒ two-tailed test

  (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.

  → let p= true proportion of employees planning to take a wellness class

  → H_0: $p = 0.42$

  H_A: $p > 0.42$ ⇒ right-tailed test

  (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.

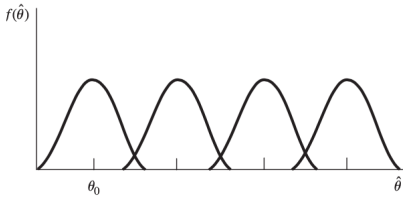
  → Let p= true proportion of votes for this candidate

  → H_0: $p = 0.5$

  H_A: $p > 0.5$ ⇒ right-tailed test

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

  e.g) $p \uparrow$ & $\bar{x}$

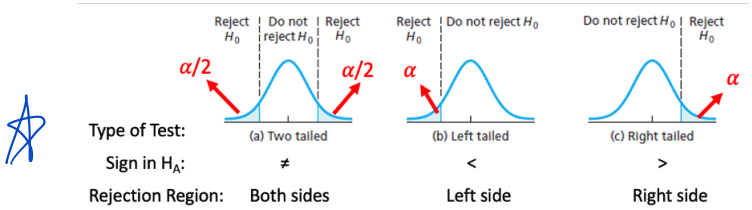


  - For example, for a right-tailed test
    - If $\hat{\theta}$ is close to θ_0, it seems reasonable to accept H_0.
    - If in reality $\theta > \theta_0$, then $\hat{\theta}$ is more likely to be large.

    Consequently, large values of $\hat{\theta}$ (relative to θ_0) favor rejection of $H_0 : \theta = \theta_0$ and acceptance of $H_A : \theta > \theta_0$.

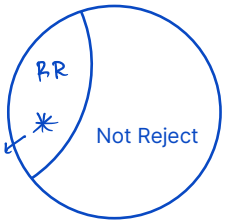
– Simply stated, we have to determine when there is or is not enough evidence against the Null based on our sample data.

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

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

$\{\hat{\theta} : \hat{\theta} \leq k \quad \text{OR} \quad \hat{\theta} \geq k\}$

– 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 $RR = \{\bar{x} : \bar{x} \geq 53\}$.

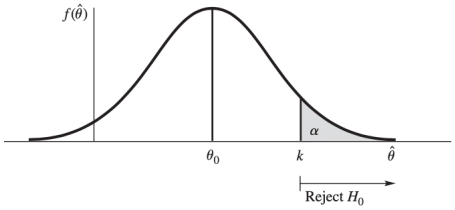
– But how did we choose the value of k? More generally, how can we find some objective criteria for deciding which value of k specifies a good rejection region of the form $\{\bar{x} \geq k\}$?

- Significance level

– The **significance level** α 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.

$100(1-\alpha)\%$ CI $= \hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}$ $\Rightarrow$ use correct θ & $\hat{\theta}$ for problem

- 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 σ_1^2 and σ_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 θ 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 θ and standard error $\sigma_{\hat{\theta}}$.

If n is large $\Rightarrow \hat{\theta} \overset{\text{approx}}{\sim} N(\theta, \sigma_{\hat{\theta}})$

$\Rightarrow \dfrac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \overset{\text{approx}}{\approx} Z \overset{\text{approx}}{\sim} N(0,1)$

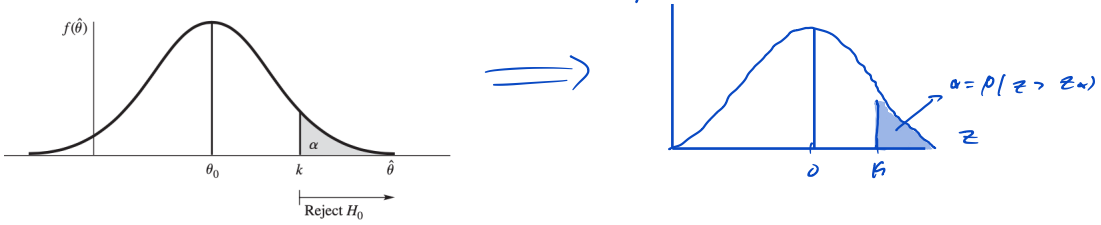
θ	$\hat{\theta}$	$\sigma_{\hat{\theta}}$
μ	$\bar{X}$	$\dfrac{\sigma}{\sqrt{n}}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
p	$\hat{p}$	$\sqrt{\dfrac{p(1-p)}{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

estimate σ_i^2 with s_i^2

* will adjust slightly when testing

- Then we have the following:

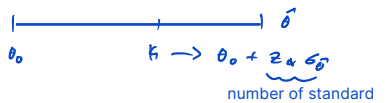
$$H_0 : \theta = \theta_0$$
$$H_A : \theta > \theta_0$$
$$TS : \hat{\theta}$$
$$RR : \{\hat{\theta} > k\}$$



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

  Assuming H_0 is true, if we desire an α-level test, then

given $H_0 \rightarrow \hat{\theta} \overset{\text{approx}}{\sim} \text{Normal}(\theta_0, \sigma_{\hat{\theta}})$ $\longrightarrow$ $RR = \{\hat{\theta} : \hat{\theta} > \theta_0 + z_\alpha \sigma_{\hat{\theta}}\}$



$$= \left\{\hat{\theta} : \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > z_\alpha\right\}$$

$$= \{z : z > z_\alpha\}$$

$$\rightarrow \hat{\theta} = \theta_0 + z_\alpha \sigma_{\hat{\theta}} \Longrightarrow \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} = z_\alpha$$

Standardize

- Thus, an equivalent form of the test, with level α is:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta > \theta_0$$
$$TS : z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$
$$RR : \{z > z_\alpha\}$$

- 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

$$Z = \frac{\overset{\text{point estimate}}{PE} - \overset{\text{null}}{H_0}}{\underset{\text{Standard error}}{SE}}$$

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).

→ feel free to shorthand, just get the main points

- Examples

one mean μ
$n = 52$
$X \sim$ Normal
σ^2 ✓

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 $\alpha = 0.05$.

Calculator → z-test

→ Let μ= true mean honey amount with new type of pollen

→ $H_0: \mu = 55$
$H_A: \mu > 55$

p-value $= P(z > 1.622) = 0.052 \not< \alpha = 0.05$
⇒ fail to reject ✗

→ TS: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{57.25 - 55}{10/\sqrt{52}} \approx 1.622$

RR: $\{ z > z_{0.05} \} = \{ z > 1.645 \}$ → fail to reject H_0 ✗
↳ invNorm(0, 1, 0.05)

→ Conclusion: Because our TS $z = 1.622 \notin$ RR $z \geq 1.645$, we fail to reject the null hypothesis. At the 5% significance level, there is not enough evidence to conclude that the true average amount of honey produced from each hive with the new pollen type is greater than 55 mL.

one mean μ
$n = 36$
$X \sim$??
σ^2 ✗

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 $\alpha = 0.025$.

Calculator → z-test

→ Let μ= true mean number of contacts made per week

→ $H_0: \mu = 15$
$H_A: \mu > 15$

p-value $= P(z > 4) \approx 0 \leq \alpha = 0.025$ ✓
⇒ Reject ✓

→ TS: $z \approx \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{17 - 15}{3/\sqrt{36}} = 4$

RR: $\{ z > z_{0.025} \} = \{ z > 1.96 \}$ → reject H_0 ✓

→ Conclusion: Because our TS $z = 4 \in$ RR $z \geq 1.96$, we reject the null hypothesis. At the 2.5% significance level, there is sufficient evidence to conclude that the true mean number of contacts made per week is greater than 15.

one proportion p

$n = 100 \rightarrow x = 15$ ✓s, $n - x = 85$ ✓s

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 $\alpha = 0.01$.

Calculator → 1 prop Z test

→ let p = true proportion of defectives

→ $H_0: p = 0.10$

$H_A: p > 0.10$

★ use p_0 because assuming H_0 is true

→ TS: $z \approx \frac{\hat{p} - p_0}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.15 - 0.10}{\sqrt{\frac{0.1(0.9)}{100}}} = 1.667$

p-value $= P(z > 1.667) = 0.047 \leq \alpha$

⇒ fail to reject

RR: $\{z > z_{0.01}\} = \{z > 2.362\}$ → fail to reject H_0 ✗

→ Conclusion: ——— follow same format + context ———

- Here is a summary of the large-sample α-level hypothesis tests:

★

Large-Sample α-Level Hypothesis Tests

$H_0 : \theta = \theta_0$.

H_a: $\theta > \theta_0$ (upper-tail alternative); $\theta < \theta_0$ (lower-tail alternative); $\theta \neq \theta_0$ (two-tailed alternative).

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

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

- 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

two, independent means $\mu_1 - \mu_2$

$n_1 = n_2 = 50$

$x_1, x_2 \sim$??

σ_1^2, σ_2^2 ✗

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
$s_1^2 = .18$	$s_2^2 = .14$

Calculator → 2 samp Z test

→ let μ_1 = true mean reaction time for men (sec)

let μ_2 = — — — — — — women — —

→ $H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

→ TS: $z \approx \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.6 - 3.8) - 0}{\sqrt{\frac{0.18}{50} + \frac{0.14}{50}}} = -2.5$

p-value $= 2\,P(z < -2.5)$

$= 2\,(0.0062)$

$= 0.0124 \leq \alpha$

⇒ Reject ✓

RR: $\{|z| > z_{\alpha/2}\} = \{|z| > 1.645\}$ → reject H_0 ✓

→ Conclusion: Because TS ∈ RR, reject H_0. At 10% sig level, there is sufficient evidence of a difference in true mean reaction time & that it is slower for women.

two proportion $p_1 - p_2$

$n_1 = 200 \rightarrow x_1 = 18 \geq 5$ ✓, $n_1 - x_1 = 182 \geq 5$ ✓

$n_2 = 600 \rightarrow x_2 = 25 \geq 5$ ✓, $n_2 - x_2 = 575 \geq 5$ ✓

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?

→ let p_1 = true proportion of defectives in line A
let p_2 = — — — — — — — — B

Calculator → 2 prop Z test

→ $H_0: p_1 - p_2 = 0$

$H_A: p_1 - p_2 \neq 0$

Assuming equivalent ⇒ pooled

$p_1 = p_2 = p \rightarrow \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{18 + 25}{200 + 600} \simeq 0.0537$

→ TS: $z \approx \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{(0.09 - 0.0416) - 0}{\sqrt{0.0537(1-0.0537)\frac{1}{200} + \frac{1}{600}}} \simeq 2.67$

↳ p-value $= 2P(z > 2.67) = 2(0.0038) = 0.0076 \leq \alpha$

⇒ Reject ✓

RR: $\{|z| > z_{0.025}\} = \{z > 1.96\}$ → Reject H_0 ✓

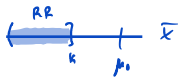
→ Conclusion: TS ∈ RR ⇒ reject H_0. At $\alpha = 0.05$, there is enough evidence to conclude the true proportion of defects is greater in line A.

Small sample tests for μ 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

→ $X \sim \text{Normal}(\mu_0, \sigma^2 = ??) \rightarrow RR = \{\bar{x} : \bar{x} < \mu_0 - t_\alpha \frac{s}{\sqrt{n}}\}$ (k)



$\Rightarrow \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = T \sim t_{n-1} \rightarrow = \{t : t < t_\alpha\}$

- Here is a summary of the small-sample α-level tests for μ

A Small-Sample Test for μ

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: $\mu > \mu_0$ (upper-tail alternative).
$\mu < \mu_0$ (lower-tail alternative).
$\mu \neq \mu_0$ (two-tailed alternative).

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

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

- If we are testing two independent means $\mu_1 - \mu_2$ and assume both Normal distributions with common unknown variance σ^2, then we use the pooled variance S_p^2 as the estimator for σ^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: $T = \dfrac{\bar{Y}_1 - \bar{Y}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, 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

$\mu_1 - \mu_2 = \mu_D \longrightarrow \dfrac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = T \sim t_{n-1} \Rightarrow$ one mean test on differences

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

- Examples

one mean μ
n = 9
X ~ normal
σ? X

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 $\alpha = 0.10$.

   Calculator → T-test

→ let μ = true mean growth of tumor with enzyme

→ $H_0: \mu = 4$

$H_A: \mu \neq 4$

→ TS: $T = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{4.2824 - 4}{1.2/\sqrt{9}} = 0.706$

p-value $= 2P(t_8 > 0.706)$
$= 1 - (0.7501)$
$= 0.5002 \not< \alpha = 0.10$
⇒ fail to reject

RR: $\{|t| > t_{0.05,8}\} = \{|t| > 1.860\}$ → fail to reject H_0 ✗
↳ invT(0.05, 8)

→ At $\alpha = 0.10$, not sufficient evidence to conclude the true mean tumor growth is not 4 mm under new enzyme

two, independent means $\mu_1 - \mu_2$
$n_1 = n_2 = 9$
$X_1, X_2 \sim ?$ ⇒ look at histograms
σ_1^2, σ_2^2 X

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
$s_1^2 = 24.445$	$s_2^2 = 20.0275$

Calculator → 2 samp Ttest

→ let μ_1 = true mean assembly time for standard procedure
let μ_2 = — — — — — — new — — — —

→ H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$

check if can pool : $\frac{s_1^2}{s_2^2} = \frac{24.445}{20.0275} \approx 1.22 \leq 2$ ✓

$\Rightarrow s_p = \sqrt{\frac{8(24.445) + 8(20.0275)}{9+9-2}}$

$\downarrow = 4.716$

→ TS: $T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(35.22 - 31.56) - 0}{4.716\sqrt{\frac{1}{9} + \frac{1}{9}}} \approx 1.646$

RR: $\{|t| > t_{\alpha/2, \nu}\} = \{|t| > 2.12\}$ → Fail to reject H_0 X
(ν: $n_1 + n_2 - 2$)

p-value $= 2\,P(t_{16} > 1.646)$
$= 2\,(0.059)$
$= 0.1193 \not< \alpha = 0.05$

⇒ fail to reject

→ Conclusion: Because TS $\notin$ RR, fail to reject H_0.
At 5% sig level, there is insufficient evidence of a difference in true mean assembly time between the standard & new procedure.

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 fail to reject, but were "closer" to rejecting H_0 example 1. 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.

$P($ more extreme TS $|$ $H_0)$

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

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

– 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 θ_0 was true, the result we got had such a tiny probability to occur. So the original assumption of θ_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**.

Reject if p-value $\leq \alpha$

fail to reject if p-value $> \alpha$

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

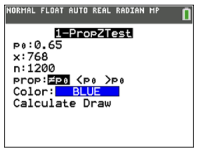
Treating it like this leaves it up to the reader to evaluate the extent to which the observed data disagree with the null hypothesis and make their own choice in α in deciding whether or not to reject H_0. This is one advantage of p-values, and is why most scientific journals require p-values for all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

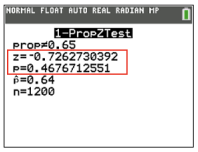
(often $\alpha = 0.1, 0.05, 0.01$ are chosen out of convenience rather than a well-thought out choice.)

Calculator session

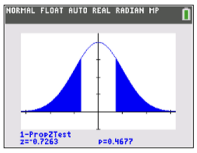
STAT > TESTS >

One proportion test: p





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

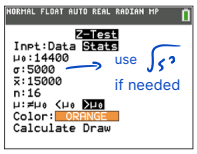


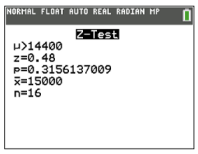
Draw Output
Plot (and displays values) of p = p-value and z = TS on the standard normal curve

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

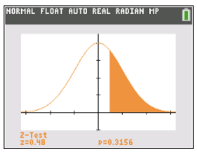
*This does NOT give us the Critical Value $z_{\alpha/2}$, we have to figure that out ourselves

One mean test: μ





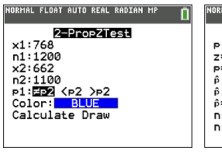
Calculate Output
μ= Alternative hypothesis
z = TS
p = p-value
$\bar{x}$= sample mean
n = sample size

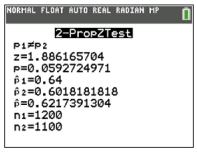


H_0: $\mu = 14{,}400$
H_A: $\mu > 14{,}400$

Two proportions test: $p_1 - p_2$

STAT > TESTS >





$H_0: p_1 - p_2 = 0$

$H_A: p_1 - p_2 \neq 0$

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 θ $\qquad$ α-level test $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$

  $\hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}$ $\qquad$ $z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \rightarrow RR = \{|z| > z_{\alpha/2}\}$

same

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

  $AR = RR^c = \left\{ -z_{\alpha/2} \leq \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2} \right\} \Leftrightarrow \hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}} \leq \theta_0 \leq \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}$

  RR AR RR

- Thus, we "accept" $H_0 : \theta = \theta_0$ if θ_0 falls within the $100(1-\alpha)\%$ CI and reject if outside.

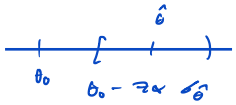
  So the confidence interval can be thought of as the set of values of θ_0 for which $H_0 : \theta = \theta_0$ is "acceptable" at level α.

  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

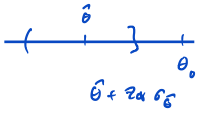
  For $H_0 : \theta = \theta_0$ with level α and $100(1-\alpha)\%$ CIs

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

  lower bound confidence interval

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

  upper bound confidence interval

Decision	Truth: H_0 TRUE	Truth: H_0 FALSE
reject H_0	Type I, α	✓ correct, $1-\beta$ = power
fail 2 reject H_0	✓ Correct	Type II, β

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. → Incorrectly rejecting H_0, false positive

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

    $\alpha = P(\text{Type I error}) = P(\text{Reject when } H_0 \text{ is TRUE}) = P(TS \in RR \mid H_0)$

  - **Type II error** is made if H_0 is accepted when H_A is true. → Incorrectly accepting H_0, false negative

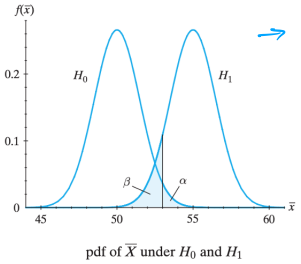
    The probability of a type II error is denoted by β.

    $\beta = P(\text{Type II error}) = P(\text{Accept } H_0 \text{ when } H_0 \text{ is FALSE}) = P(TS \notin RR \mid H_A)$

- We can think of α and β 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 $H_A : \mu = 55$), $n = 16$, $\sigma^2 = 36$ $\Rightarrow \bar{x} \sim \text{Normal}(\mu, \sigma^2 = \frac{36}{16})$



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

→ Suppose $RR = \{\bar{x} > 53\} \Rightarrow \alpha = P(\bar{x} \geq 53 \mid \mu = 50)$ (H_0 True)

$\downarrow = 0.0228$

$\beta = P(\bar{x} < 53 \mid \mu = 55)$ (H_A True)

$\downarrow = 0.0912$

$\Rightarrow$ Power $= 1 - \beta = 1 - 0.0912$

$\downarrow = 0.9088$

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

  The larger the difference is between θ 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 $H_A : \mu = 51$.

  $\beta = P(\bar{x} < 53 \mid \mu = 57)$ $\quad$ $\beta = P(\bar{x} < 53 \mid \mu = 51)$

  $\beta = 0.0038$ $\quad$ $\beta = 0.9088$

  Power $= 0.9962$ $\quad$ Power $= 0.0912$

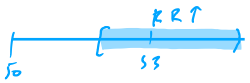
- This example shows that the test using $RR = \{\bar{x} \geq 53\}$ guarantees a low risk of making a type I error $\alpha = 0.023$, but it does not offer adequate protection against a type II error (high βs with some alternative hypotheses).

- 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. α and β are inversely related).

- So how can we improve our test? One way is to balance α and β 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.

  $\alpha \uparrow \Rightarrow \beta \downarrow$ $\quad$ ex) 

- 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 : The elevator is not safe (repairs are needed)

  H_A : The elevator is safe (no repairs needed)

  – Consequences of Type I error: Incorrectly assume elevator is safe, when in reality it is not. People could be hurt. $\rightarrow$ worse $\Rightarrow$ minimize α

  – Consequences of Type II error: Incorrectly assume the elevator needs repairs, when in reality it does not. Spend money on unnecessary repairs.

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

  Intuitively obvious, collect more data!