Lecture 22: Hypothesis Testing

MATH 697

Omidali Jazi

November 15, 2018

McGill University

Goals for this Chapter

- Introduction to Testing Hypotheses
  - Steps in Conducting a Hypothesis Test
  - Types of Errors
  - Some terminologies
- Large Sample Tests for the Means and Proportions
  - Z tests
  - Type II Error Probability Calculations
  - Sample Size Calculations
- The p-value approach to hypothesis testing
- Small-sample hypothesis tests (means, proportions)
- Tests for variance(s)
- Neyman-Pearson Lemma and likelihood ratio tests

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 or as tests of hypotheses about their values.
- Scientists pose a hypothesis concerning one or more population parameters that they equal specified values. They then samples the population and compares her observations with the hypothesis.
- If the observations disagree with the hypothesis, the hupothesis will be rejected. If not, the scientists 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.

Example 1

- 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. The new drug A is given to the patients in the first group, and the old drug B is given to the patients in the second group.
- 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.

Example 2

- You hypothesize a coin is fair
  – To test, take a coin and start flipping it
  – If it is fair, you expect that about half of the flips will be heads and half tails
  – After a large number of flips, if you see either a large fraction of heads or a large fraction of tails you tend to disbelieve your hypothesis
- This is an informal hypothesis test.
- How to decide when the fraction is too large?

Is the Coin Fair?

- Example: Flip a coin 10 times and count the number of heads.
- Assuming the coin is fair (and flips are independent), the number of heads has a binomial distribution with $n = 10$ and $p = 0.5$
- So, if our assumption is true, the distribution of the number of heads is:

Number of heads (x):	0	1	2	3	4	5	6	7	8	9	10
Pr(X=x) =	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

- And, if your assumption is not true, what do we expect to see? Either too few or too many heads.

Setting Up a Test for the Coin

- Idea: Make a rule, based on the number of heads observed, from which we will conclude either that our assumption ($p = 0.5$) is true or false.
- Here is one rule: Assume that $p = 0.5$ if $3 \leq x \leq 7$
  – Otherwise, we conclude that $p \neq 0.5$

- If p=0.5, what's our chance of making a mistake?

Number of heads (x):	0	1	2	3	4	5	6	7	8	9	10
Pr(X=x) =	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

~11%

- If p=0.8, what's our chance of detecting the biased coin?

Number of heads (x):	0	1	2	3	4	5	6	7	8	9	10
Pr(X=x) =	0.000	0.000	0.000	0.001	0.006	0.026	0.088	0.201	0.302	0.268	0.107

~68%

Cofidence Interval vs. Hypothesis Testing

- Previously we would have answered this question with a confidence interval:
  – Say we observed $y/n = 0.7$: we are 95% confident that the interval $[0.5, 0.9]$ covers the true p.
- Now, we look to answer the question using a hypothesis test:
  – If the true probability of heads is $p = 0.5$ (i.e., the coin is fair), how unlikely would it be to see 7 heads out of 10 flips?
- If we see an outcome inconsistent with our hypothesis (the coin is fair), then we reject it.

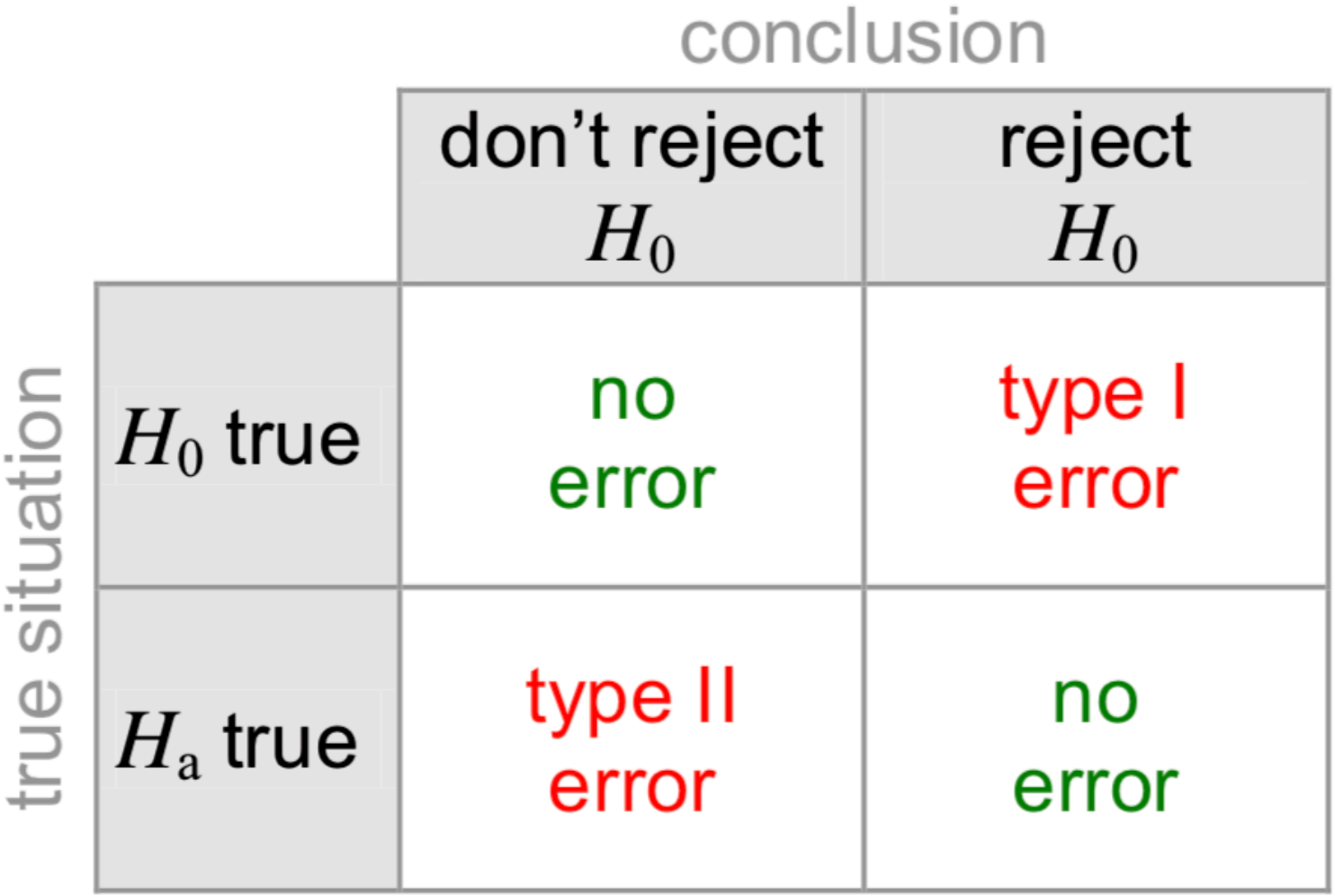
Why We Do Hypothesis Tests?

- Confidence intervals provide more information than hypothesis tests. But often we need to test a particular theory.
  – Example: A drug can lower blood pressure.
- Sometimes confidence intervals are hard/impossible
  – When the theory you are testing has several dimensions.
  – When "intervals" do not make sense.
  Example: Are two categorical variables independent?

Elements of a Statistical Test

- Null hypothesis
  - Denoted H_0
  - We believe the null unless/until it is proven false.
- Alternative hypothesis
  - Denoted H_a
  - It is what we would like to prove.
- Test statistic
  - The test statistic is the empirical evidence from the data.
- Rejection region
  - If the test statistic falls in the rejection region, we say "reject the null hypothesis" or "we have proven the alternative"
  - Otherwise, we "fail to reject the null hypothesis".

Errors in Hypothesis Testing



Probability of Type I Error

- The Probability of Type I Error which are also known as significance level or the level of the test

$$\alpha = P(\text{Type I error}) = P(H_0 \text{ is rejected when it is true})$$

- Experimenters choose prior to the test.
  - Conventions: α =0.10, 0.05, and 0.01
  - Can increase or decrease, depending on the problem.

- Choice of α defines rejection region (or size of p-value) that results in rejection of the null hypothesis.

Probability of Type II Error

- Probability of Type II Error

  $\beta = P(\text{Type II error}) = P(\text{Not rejecting } H_0 \text{ when it is false})$

- It is a function of the actual alternative distribution and the sample size.

- $1 - \beta$ is called the power of a test
  – It is $P(\text{Rejecting } H_0 \text{ when it is false})$

- Ideal is both small α and β (i.e., high power), but for a fixed sample size they trade off.
  – By convention, we control α by choice and β with sample size. (bigger sample, more power)

Choosing the Null and Alternative

- In hypothesis testing, we get to control the Type I error.
  – So, if one error is more severe than the other, set the test up so that it is the Type I error.
- Example: Sending an innocent person to jail more serious than letting a guilty person go free. Hence, the burden of proof of guilt is placed on the prosecution at trial. ("innocent until proven guilty").
- The null hypothesis is a person is innocent, and the trial process controls the chance of sending an innocent person to jail.

Example

Consider a political poll of $n = 15$ people. We want to test $H_0 : p = 0.5$ vs. $H_a : p < 0.5$, where p is the proportion of the population favoring a candidate. The test statistic is Y, the number of people in the sample favoring the candidate

- If the $RR = \{y \leq 2\}$, find the level of the test (α).
- If $p = 0.3$ what is the probability of a Type II error (β)?
- If $p = 0.1$, what is the probability of a Type II error (β)?
- If $RR = \{y \leq 5\}$
  - What is the level of the test (α)?
  - If $p = 0.3$, what is β?

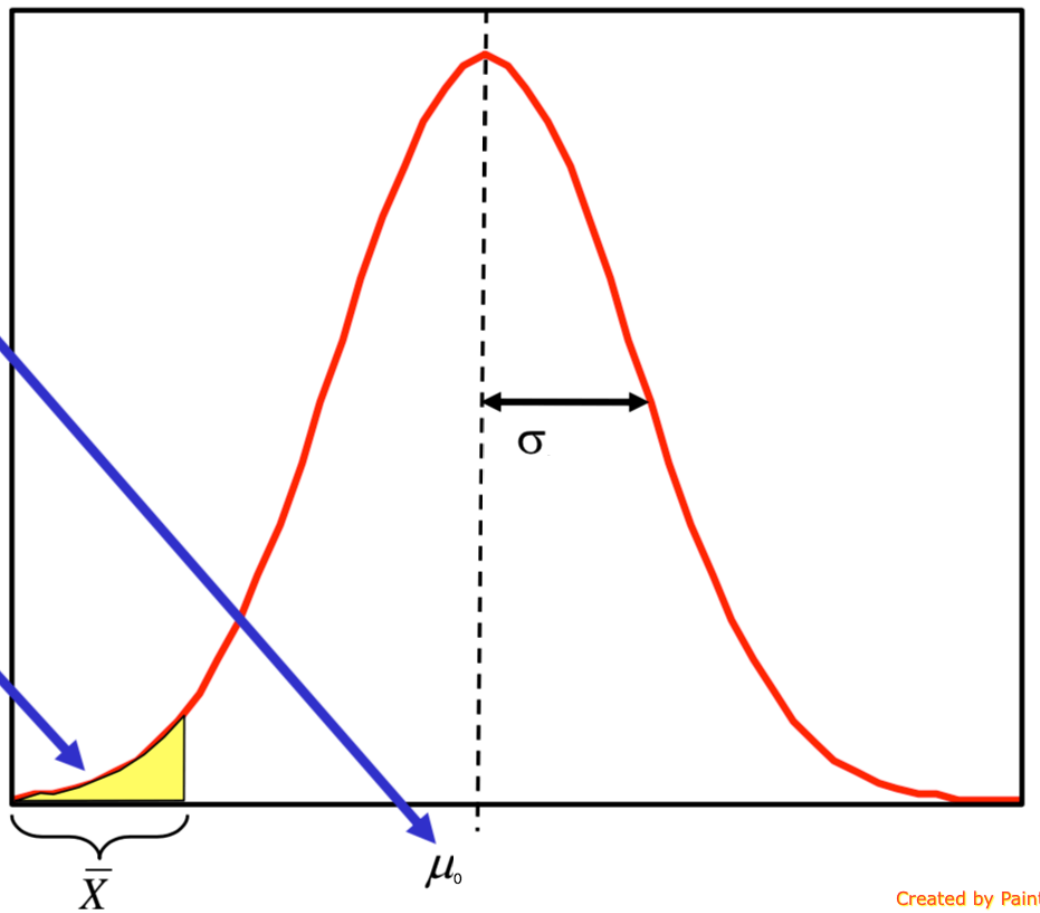
Terminology

- Null and Alternative hypotheses.
  – We will believe the null until it is proven false.
- Acceptance vs. Rejection region.
  – The null is proven false if the test statistic falls in the rejection region.
- Type I error vs. Type II error
  – Type I: Rejecting H_0 when it is really true.
  – Type II: Not Rejecting H_0 when it is really false.
- Significance level or level of the test (α)
  – Probability of the Type I error.
- $P(\text{Type II error}) = \beta$ and $1 - \beta$ is called the power of the test.
  – It is a function of the actual alternative distribution.

Intuition Behind the Test ($H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$)

1. We will assume the status quo is true…
   a. In this case, that there is no change in the mean

2. …unless there is sufficient evidence to contradict it
   a. In this case, meaning we observe a much smaller sample mean than we would expect to see assuming μ_0



Steps in Hypothesis Testing

1. Identify the parameter of interest
2. State null and alternative hypotheses
3. Determine form of test statistic
4. Calculate rejection region
5. Calculate test statistic
6. Determine test outcome by comparing test statistic to rejection region

Identify Parameter of Interest

- In hypothesis tests, we are testing the parameter of a distribution.
  – Example: μ or σ for a normal distribution.
  – Example: p for a binomial distribution.
  – Example: α and β for a gamma distribution.

- So, the first step is to identify the parameter of interest.
  – Often we will be testing the mean μ of a normal, since the CLT often applies to the sample mean.
  - Example: For the coin and election examples, we are testing p.

State the Null and Alternative Hypothesis

- H_0: The null hypothesis is a specific theory about the population that we wish to disprove.
  – We will believe the null until it is disproved.
- In notation, $H_0 : \mu = \mu_0$
- H_a: The alternative hypothesis is what we want to prove
  – What we will believe if the null is rejected.
- In notation, $H_a : \mu > \mu_0$, $H_a : \mu < \mu_0$, or $H_a : \mu \neq \mu_0$

Null vs Alternative Hypotheses

- The null hypothesis is what you have assumed.
  – Generally, it is the status quo or less desirable test outcome.
  – Failing to prove the alternative does not mean the null is true, only that you do not have enough evidence to reject it.
- The alternative is proven based on empirical evidence
  – It is the desired test outcome and/or the outcome upon which the burden of proof rests
  – The significance level (α) is set so that the chance of incorrectly "proving" the alternative is tolerably low.
- Having proved the alternative is a much stronger outcome than failing to reject the null.
  – Thus, structure the test so the alternative is what needs proving.

Example

- In the example, we want to test is whether the drug will lower blood pressure.
  – So, the null hypothesis is the status quo and the alternative carries the burden of proof to show otherwise.
  – We write this out as $H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$
- The other possibilities are $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$ and $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$
  – What would you be testing with these hypotheses?

Expressing the Null as an Equality

- We will express the null hypothesis as an equality and the alternative as an inequality.
  - Example: $H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$
- In reality, the hypotheses divide the real line into two separate regions.
  - Example: $H_0 : \mu \geq \mu_0$ vs $H_a : \mu < \mu_0$
- However, the most powerful test occurs when the null hypothesis is at the boundary of its region. Hence, we write the null as an equality.

Determine the Test Statistic (and its Sampling Distribution)

- The test statistic is a function of the sample statistic corresponding to population parameter you are testing.
- Population and sample statistic examples:
  - Population mean $\Leftrightarrow$ Sample mean
  - Difference of two population means $\Leftrightarrow$ Difference of two sample means.
- It is sometimes "a function of" the sample statistic as we may rescale the sample statistic.
- We use the sampling distribution to determine whether the observed statistic is "unusual".

Example

- In the example, we are testing the population mean μ, so the obvious choice for the sample statistic is $\bar{X}$.
- In this case, it is easier to make the test statistic the rescaled sample statistic,

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

  where μ_0 is the null hypothesis mean.
- Why? Because, assuming the null hypothesis is true, we know the sampling distribution of Z is $N(0, 1)$
  – This is exactly true if X has a normal distribution and approximately true via the CLT for large sample sizes.

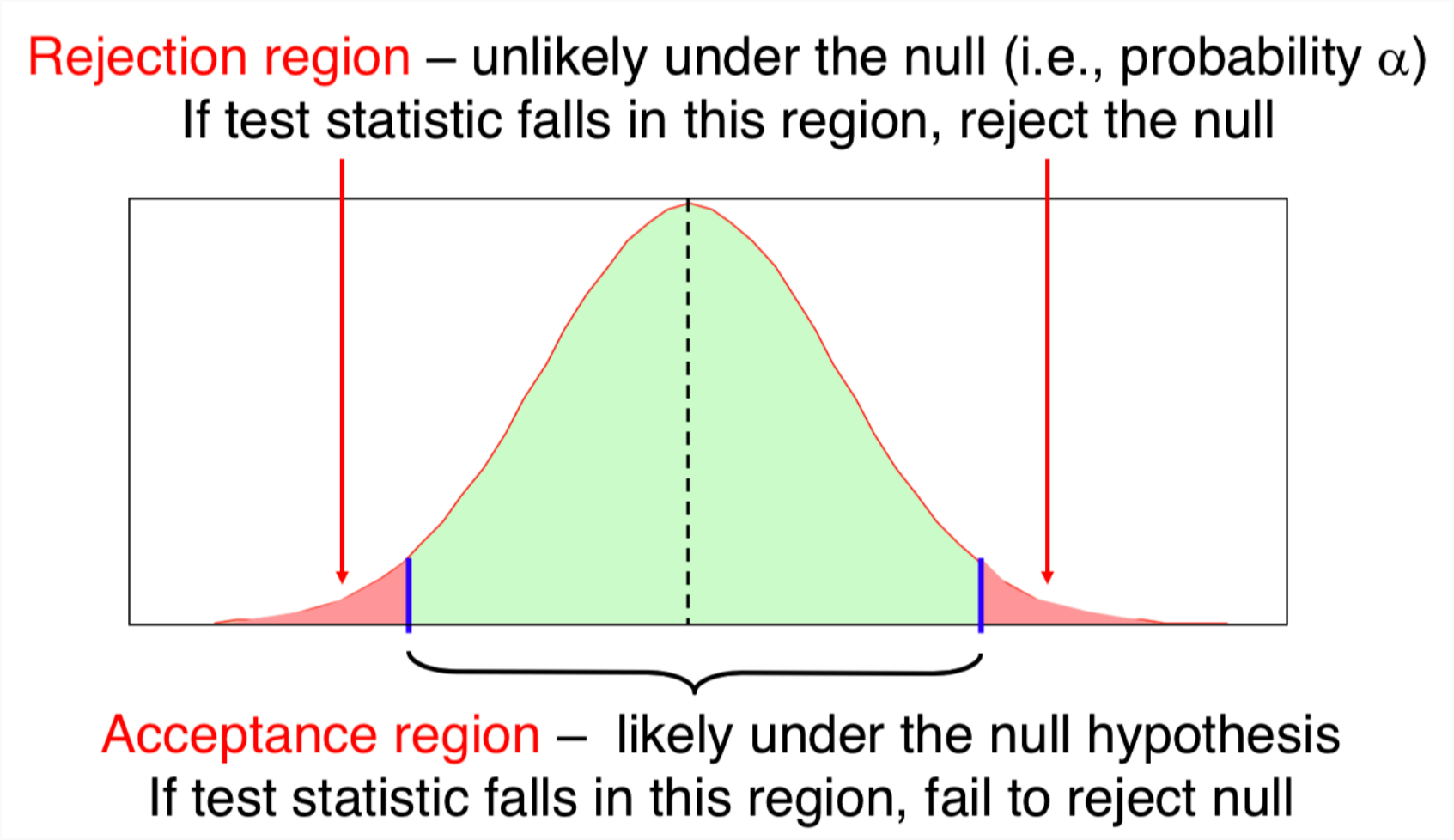
Calculate Rejection Region

- Rejection region depends on the alternative hypothesis
- Set the significance level α so that the

  $\alpha = P(\text{Fall in the rejection region} | \text{ The null Hypothesis is true})$

- This means that you will have to determine the appropriate quantile or quantiles of the sampling distribution.

The Way to Think About It (for a "two-sided" test)



Calculate the Test Statistic and Determine the Outcome

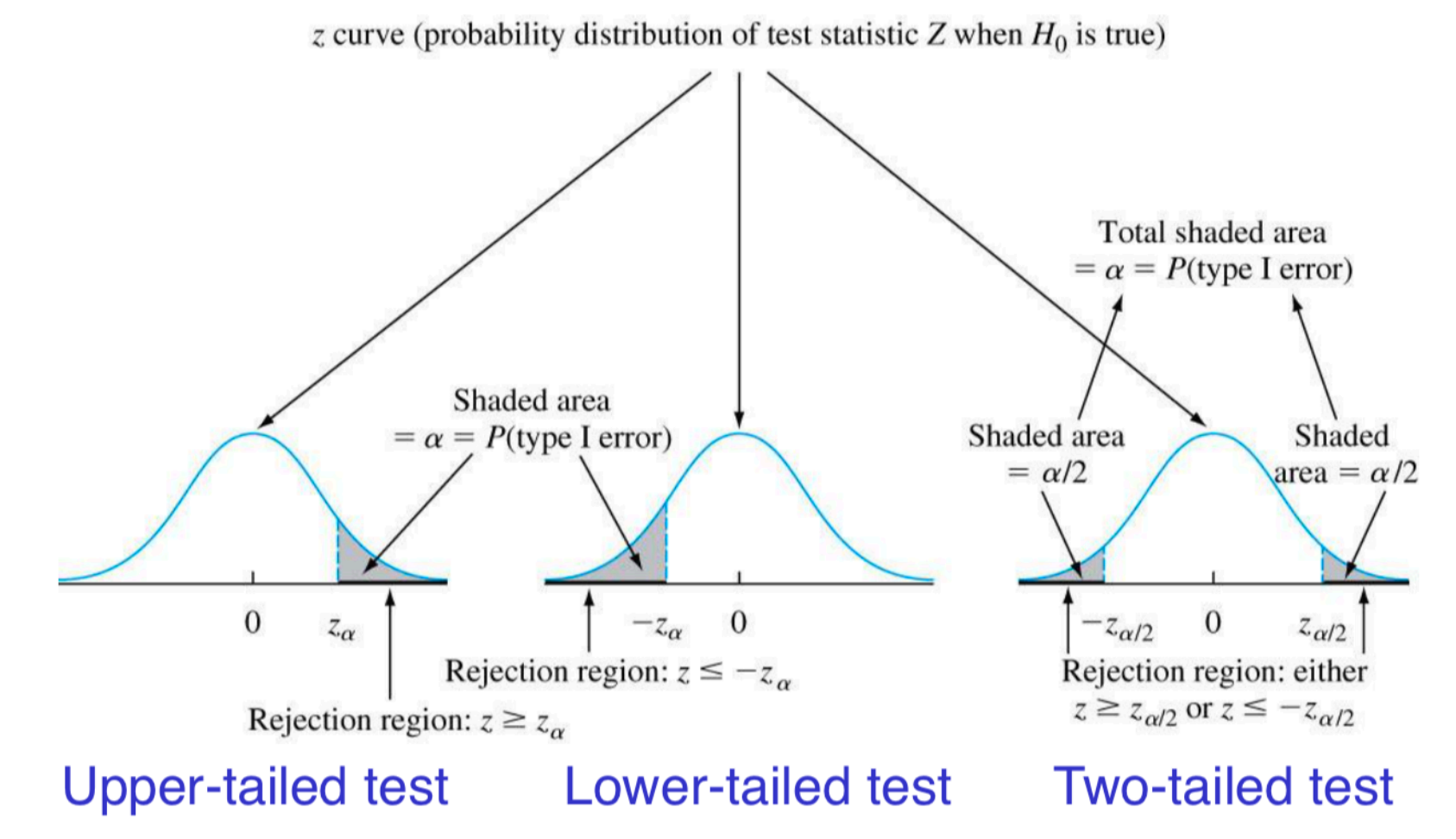
- Now, plug the necessary quantities into the formula and calculate the test statistic
- Compare the test statistic with the rejection region.
- If it falls within the rejection region you have rejected the null hypothesis.
  - Equivalently, "proven the alternative".
- If it falls in the acceptance region, you have failed to reject the null hypothesis.
  - Equivalently, "failed to prove the alternative"

Large-Sample or z-Tests

- The statistic is (approximately) normally distributed
- The rescaled test statistic is $Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$
- The null hypothesis is $H_0 : \theta = \theta_0$
- Three possible alternative hypotheses and tests:

Alternative Hypothesis	Rejection Region for Level α Test	
$H_a: \theta > \theta_0$	$z \geq z_\alpha$	(upper-tailed test)
$H_a: \theta < \theta_0$	$z \leq -z_\alpha$	(lower-tailed test)
$H_a: \theta \neq \theta_0$	$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$	(two-tailed test)

Picturing z-Tests



Example

- VP claims mean contacts/week is more than 15. Data collected on random sample of $n = 36$ people. Given $\bar{y} = 17$ and $s^2 = 9$, is there evidence to refute the claim at a significance level of $\alpha = 0.05$?
- Solution: The test is $H_0 : \mu = 15$ vs $H_a : \mu < 15$The test statistic is $Z = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$ which is approximately normal because of the CLT. So, $Z = \frac{2}{3/\sqrt{36}} = 4$ which is larger than $z_{0.05} = 1.645$. Hence, there is enough evidence to reject H_0. The mena contact per week is more than 15.

Large Sample Tests for Population Proportion (p)

- If we have a large sample, then via the CLT, $\hat{p} = y/n$ has an approximate normal distribution.
- For the null hypothesis $H_0 : p = p_0$, there are three possible alternative hypotheses

Alternative Hypothesis	Rejection Region for Level α Test	
$H_a: p > p_0$	$z \geq z_\alpha$	(upper-tailed test)
$H_a: p < p_0$	$z \leq -z_\alpha$	(lower-tailed test)
$H_a: p \neq p_0$	$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$	(two-tailed test)

where $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$ is the test statistic.

What is a Large Sample test?

- Remember that the y in $\hat{p} = y/n$ is a sum of bernoulli variables.
  – If n is sufficiently large, then the CLT "kicks in" and you can assume $\hat{p}$ has an approximately normal distribution
- Rule of thumb for using the CLT: $p \pm 3\sqrt{p_0 q_0/n}$ lies in the interval $(0, 1)$ or equivalently,

$$n > 9\left(\frac{\max(p_0, q_0)}{\min(p_0, q_0)}\right)$$

Example

- Machine must be repaired is produces more than 10% defectives. A random sample of $n = 100$ items has 15 defectives. Is there evidence to support the claim that the machine needs repairing at a significance level of $\alpha = 0.01$?

  Solution: The test is $H_0 : p = 0.1$ vs $H_a : p > 0.1$. The test statistic is equal to $z = \frac{0.15-0.10}{\sqrt{0.10\times 0.90/100}} = \frac{5}{3} = 1.667$. Comparing to $z_{0.01} = 2.326$, we conclude that there is not enough evidence to reject H_0.

Calculating Type II Error Probabilities

- The probability of a Type II error (β) is the probability a test fails to reject the null when the alternative hypothesis is true.
  - Note that it depends on a particular alternative hypothesis.
- We can write it mathematically as

$$P(\text{Not rejecting } H_0 | H_a \text{ is true with } \theta = \theta_a)$$

  - To determine, first figure out the rejection region (a function of the null hypothesis), then calculate the probability of falling in the acceptance region when $\theta = \theta_a$.

Calculating Type II Error Probabilities

- Example: Consider the test $H_0 : \theta = \theta_0$ vs $H_a : \theta > \theta_0$. Then the rejection region is of the form $RR = \{\hat{\theta} : \hat{\theta} > k\}$ for some value of k.
- So, the probability of a Type II error is

$$\begin{aligned}\beta(\theta_a) &= P(\hat{\theta} \text{ is not in RR when } H_0 \text{ is false}) \\ &= P(\hat{\theta} \leq k | \theta = \theta_a \text{ where } \theta_a > \theta_0) \\ &= P(\frac{\hat{\theta} - \theta_a}{\sigma_{\hat{\theta}}} \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}})\end{aligned}$$

Alternatively, the Power of a Test

- Recall that the power of a test is the probability a test will reject the null correcttly, for a particular alternative hypothesis $(1 - \beta)$.
- Why is this important?
  – Prior to doing a test, natural problem is that you want to make sure you have sufficient power to prove "interesting" alternatives.
  – Sometimes after a test results in a null result, you might want to know the probability of rejecting at the observed level.

Sample Size Calculations

- The sample size n for which a level α test also has β at the alternative value μ_a is

$$n = \begin{cases} \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_a - \mu_0}\right]^2 & \text{for a one-tailed test} \\ \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_a - \mu_0}\right]^2 & \text{for a two-tailed test} \end{cases}$$

- Here z_α and z_β are the quantiles of the normal distribution for α and β.

Relationship Between Hypothesis Tests and C.Is

- The large sample hypothesis test (z-test) is based on the statistic $Z = \frac{\hat{\theta}-\theta_0}{\sigma_{\hat{\theta}}}$, where the acceptance region is

$$\overline{RR} = \left\{ -z_{\alpha/2} \leq \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2} \right\}$$

which can be expressed as

$$\overline{RR} = \left\{ \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta_0 \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}} \right\}$$

Relationship Between Hypothesis Tests and C.Is

- Now, remember the general form for a two-sided large sample confidence interval:

$$100(1-\alpha)\%\ C.I = [\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$$

- Note the similarities to the acceptance region:

$$\overline{RR} = \left\{\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta_0 \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right\}$$

- So, if $\theta_0 \in [\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$, we would fail to reject the hypothesis test.

  - We can interpret $100(1-\alpha)\%$ C.I as the set of all values for θ_0 for which $H_0 : \theta = \theta_0$ is "acceptable" at level α.

An Alternate Approach to Hypothesis Testing: p-Values

- Current hypothesis testing approach: Specify α and report whether null is rejected.
  – Turn the results into a simple binary outcome.
  – It leaves out some potentially useful information since it does not say how strongly or weakly null was rejected (or not)
- The p-value approach provides this information.
  – It is also the standard method for hypothesis testing in all stats packages

What is a p-value?

- A p-value is the probability of seeing a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true.
  – "As extreme or more extreme" is defined by the alternative hypothesis
- The idea: Rather than checking whether the test statistic falls in the rejection region, we assess how unusual our test statistic is with a p-value.
  – If this probability is smaller than a pre-specified significance level a (usually, α= 0.1. 0.05, or 0.01), reject the null.
  – The smaller the p-value, the less likely the null is to be true

Attained significance level

Definition 1

If W is a test statistic, the p-value or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

Interpreting p-values

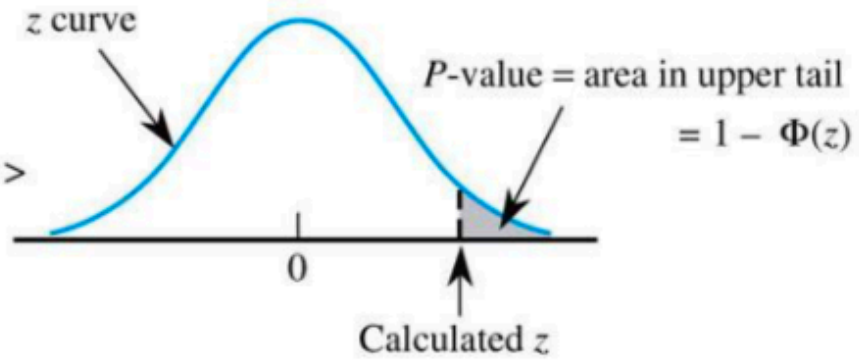
- Compare the p-value to the significance level α to decide whether to reject H_0.
- Small p-values mean either a rare event happened or the null hypothesis is false.

Why Do I Care About p-values?

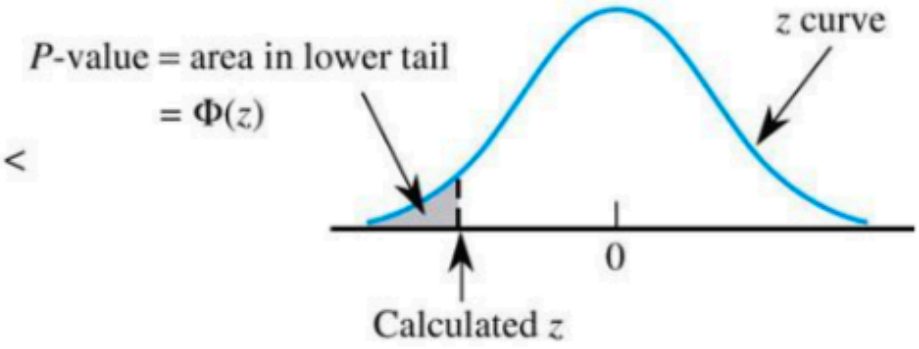
- For some types of hypothesis tests it is not possible (or hard) to calculate the acceptance/rejection region
- All stats software packages report the results of hypothesis tests in terms of p-values.
- As we have discussed, p-values provide additional information about the strength of the observed effect

Calculating p-values for z-tests

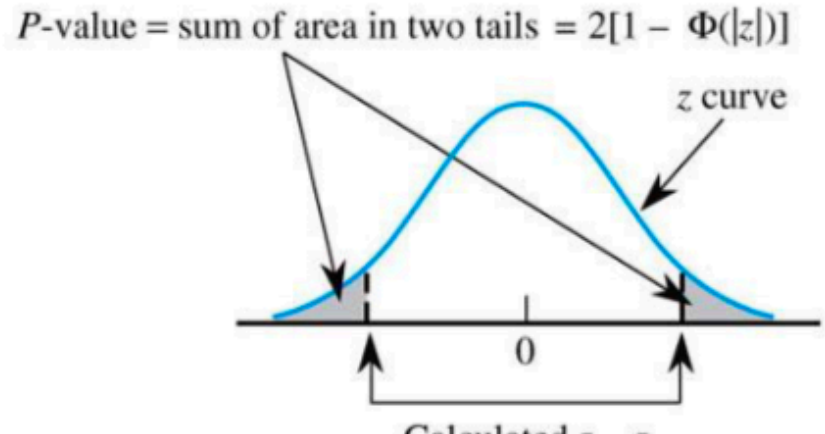
1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <



3. **Two-tailed test**
   H_a contains the inequality ≠



Steps in Hypothesis Testing When Using p-values

1. Identify the parameter of interest
2. State null and alternative hypotheses
3. Determine form of test statistic and its sampling distribution
4. Calculate the p-value
5. Determine test outcome by comparing the p-value to the significance level

Example

- You are an analyst at a logistics depot and you have been asked to determine if the average monthly of demand of an end-item is what was originally planned: 245 units per month.
  – For a sample of size $n = 50$ months, you calculate $\bar{y} = 246.18$ and $s = 3.6$
- Using the p-value approach, conduct a hypothesis test of $H_0 : \mu = 245$ vs $H_a : \mu \neq 245$, at significance level $\alpha = 0.01$.

Small-Sample Hypothesis Testing for the Mean: t-Tests

- If $Y_1, \ldots, Y_n$ has a normal distribution with unknown mean μ and unknown variance σ^2, then for the null hypothesis $H_0 : \mu = \mu_0$ the test statistic has t-distribution with $n - 1$ degrees of freedom:

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

- As always, three possible hypotheses and tests:

Alternative Hypothesis	Rejection Region for Level α Test	
$H_a: \mu > \mu_0$	$T \geq t_{\alpha,n-1}$	(upper-tailed test)
$H_a: \mu < \mu_0$	$T \leq -t_{\alpha,n-1}$	(lower-tailed test)
$H_a: \mu \neq \mu_0$	$T \geq t_{\alpha/2,n-1}$ or $T \leq -t_{\alpha/2,n-1}$	(two-tailed test)

Small-Sample Hypothesis Testing for Differences in Means: t-Tests

- For two samples, $Y_1, ..., Y_{n_1}$ and $Y_1, ..., Y_{n_2}$, from two independent population both normally distributed with unknown means and unknown but equal variances, the standardized statistic has t-distribution with $n_1 + n_2 - 2$ degrees of freedom:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - D_0}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- With $H_0 : \mu_1 - \mu_2 = D_0$ three possible tests:

Alternative Hypothesis	Rejection Region for Level α Test
$H_a : \mu_1 - \mu_2 > D_0$	$T \geq t_{\alpha, n_1+n_2-2}$ (Upper-tailed test)
$H_a : \mu_1 - \mu_2 < D_0$	$T \leq -t_{\alpha, n_1+n_2-2}$ (Lower-tailed test)
$H_a : \mu_1 - \mu_2 \neq D_0$	$T \geq t_{\alpha/2, n_1+n_2-2}$ or $T \leq -t_{\alpha/2, n_1+n_2-2}$ (Two-tailed test)

Small-Sample Tests for Population Proportion (p)

- The large-sample method is approximating the sampling distribution of $\hat{p}$.
- If the sample is small, the approximation no longer holds and we must use the actual sampling distribution.
- Rather than deriving the sampling distribution of $\hat{p}$. let's make the test statistic Y.
  - We know its distribution
  - It is binomial.
  - We are back to our original coin flipping problem.

How to Implement the Test?

- Proceed much like we did at the start of the last lecture, but now only need to determine the p-value based on the observed count.
  – As before, first define the hypotheses and specify the significance level α

- Example 1: Consider a lower-tailed test where the rejection region is defined as $Y \leq c$.
  – Then p-value $= P(Y \leq y | Y \sim Binom(n, p_0))$.

- Example 2: Consider an upper-tailed test where the rejection region is defined as $Y \geq c$.
  - Then p-value $= P(Y \geq y | Y \sim Binom(n, p_0))$.

Example

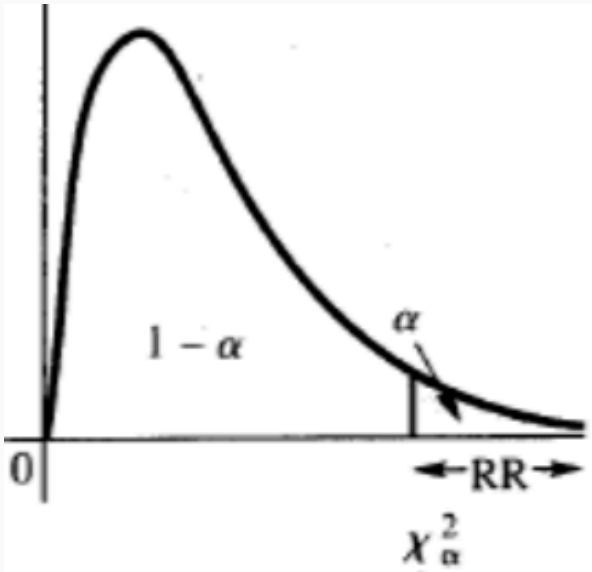
Calculate the p-value for a small-sample test of $H_0 : p = 0.9$ vs $H_a : p < 0.9$ if we observe $y = 14$ successes out of $n = 20$ trials. What do you conclude for a level $\alpha = 0.05$ test?

Testing the Variance

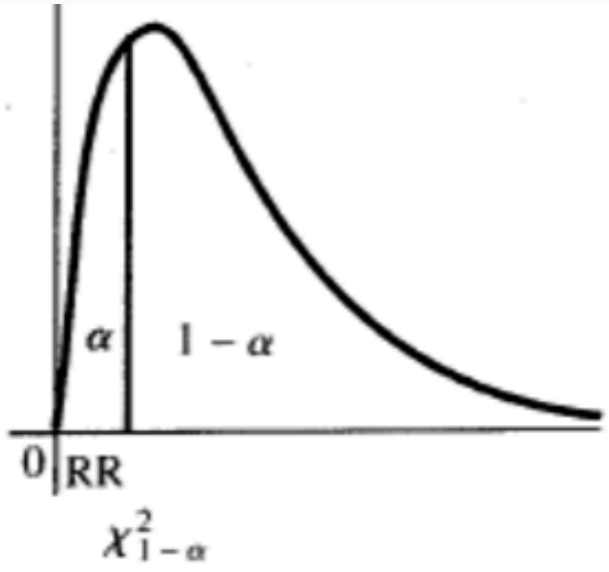
- If $Y_1, \ldots, Y_n$ have a normal distribution with unknown mean μ and unknown variance σ^2
- Null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
- Test statistic: $\mathcal{X}^2 = (n-1)S^2/\sigma_0^2$
- Three possible alternative hypotheses:

Alternative Hypothesis	Rejection Region for Level α Test
$H_a: \sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha,n-1}$
$H_a: \sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha,n-1}$
$H_a: \sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2,n-1}$ or $\chi^2 < \chi^2_{1-\alpha/2,n-1}$

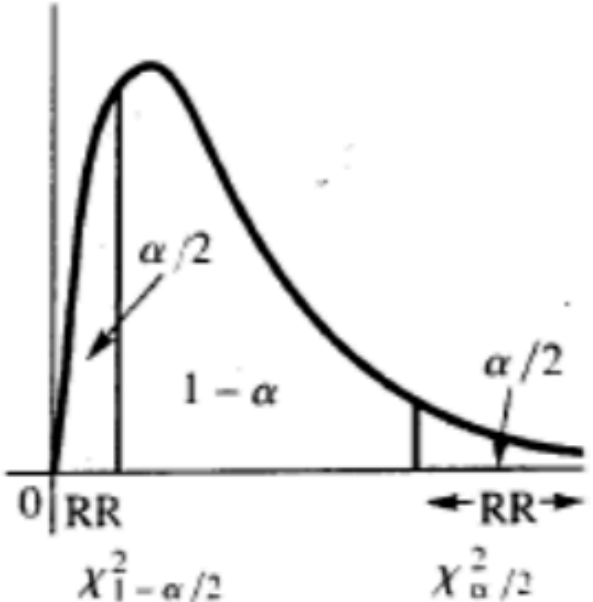
Illustrating the Rejection Region



(a)



(b)



F Distribution

Theorem 2

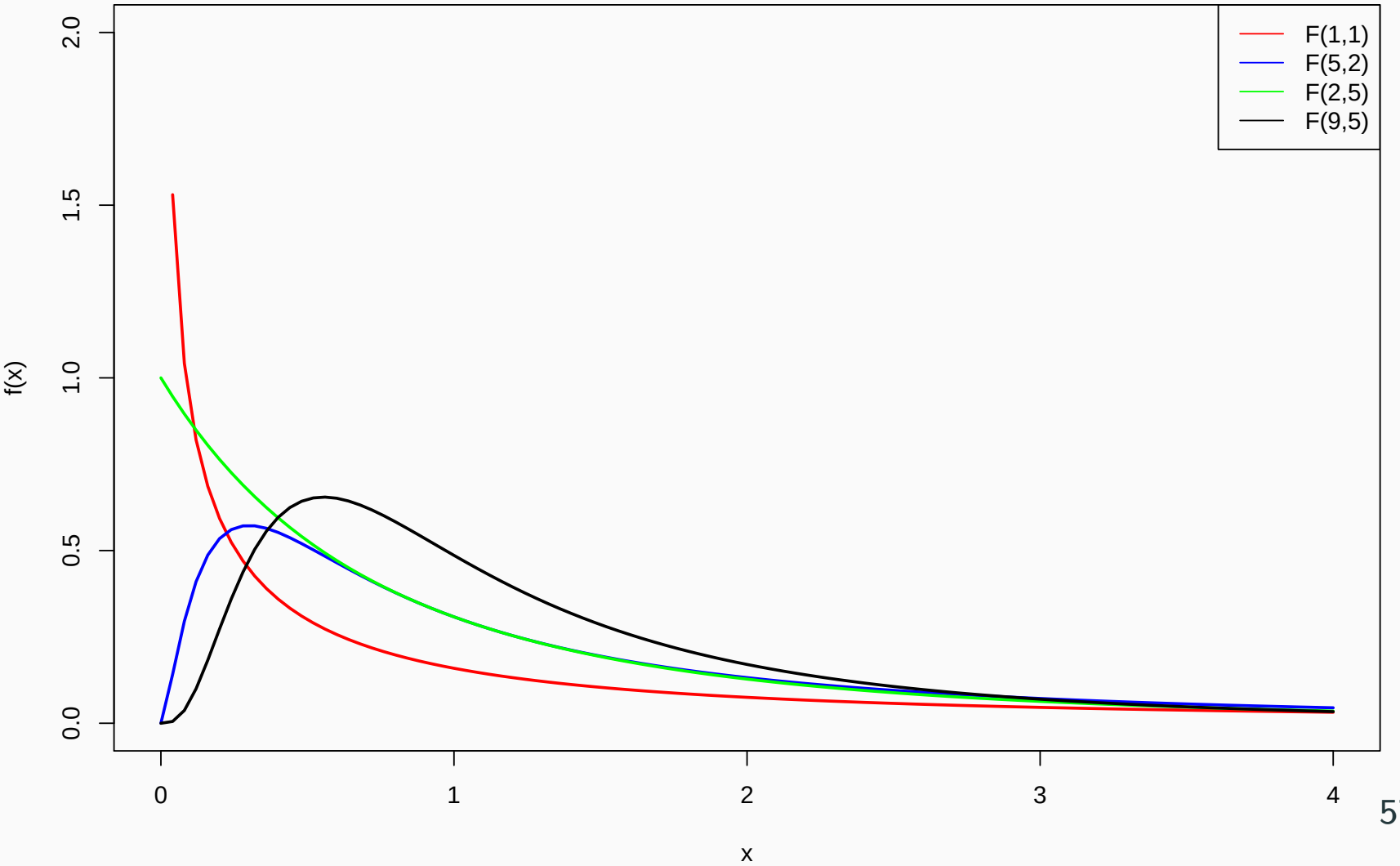
Let W_1 and W_2 be independent chi-square distributed random variables with ν_1 and ν_2 degrees of freedom, respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has F distribution with ν_1 and ν_2 degrees of freedom.

F Distribution Curves

```
curve(df(x, df1 = 9, df2  = 5), add = TRUE, col = "black",lwd = 2)
legend("topright", legend = c("F(1,1)", "F(5,2)", "F(2,5)","F(9,5)"),
       col = c("red","blue","green","black"), lty = 1)
```



F Test for Equality of Variances

- Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$
- Test statistic is $F = S_1^2/S_2^2$ where numerator has n_1 observations and denominator n_2 observations.
- Three possible alternative hypotheses:

Alternative Hypothesis	Rejection Region for Level α Test
$H_a: \sigma_1^2 > \sigma_2^2$	$F > F_{\alpha, n_1-1, n_2-1}$
$H_a: \sigma_1^2 < \sigma_2^2$	$F < F_{1-\alpha, n_1-1, n_2-1}$
$H_a: \sigma_1^2 \neq \sigma_2^2$	$F > F_{\alpha/2, n_1-1, n_2-1}$ or $F < F_{1-\alpha/2, n_1-1, n_2-1}$

Calculating the Rejection Region for a One-Tailed Test

- The easy test to conduct using the F-table is $H_0 : \sigma_1^2 > \sigma_2^2$.
  – Why? Because the table is set up to calculate rejection regions in the right tail.
  – Now note that the choice of which population is labeled "1" is arbitrary, so can always turn a left tailed test into a right tailed on.
  – This avoids having to solve for a rejection region from the Table:

$$F_{1-\alpha,\nu_1,\nu_2} = 1/F_{\alpha,\nu_2,\nu_1}$$

Calculating the Rejection Region for a Two-Tailed Test

- For $H_0 : \sigma_1^2 \neq \sigma_2^2$, we cannot avoid the left tail calculation issue.
  – Start by solving for $F_{\alpha/2,\nu_1,\nu_2}$
  - Also solve for $F_{1-\alpha/2,\nu_1,\nu_2} = 1/F_{\alpha/2,\nu_2,\nu_1}$
  - Then

$$RR = \{F : F > F_{\alpha/2,n_1-1,n_2-1} \text{ or } F < F_{1-\alpha/2,n_1-1,n_2-1}\}$$

- Example: Find the rejection region for a two-sided test with $n_1 = 5$ and $n_2 = 6$ at the significance level $\alpha = 0.05$.

Example: One-Tailed F test

```r
data1<-c(97,5,8,73,8,88,4,156,80,7,96,3,11,79,5,89,1,119,98,5)
 n1<-length(data1); #sd(data1)    #50.00463
data2<-c(25,15,18,43,16,30,2,20,75,13,30,15,21,45,5,89,1,19,50,15)
n2<-length(data2); #sd(data2)   #22.98804
var.test(data1,data2,alternative = "greater")
```

```
##
##  F test to compare two variances
##
## data:  data1 and data2
## F = 4.7317, num df = 19, denom df = 19, p-value = 0.0006954
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  2.182262      Inf
## sample estimates:
## ratio of variances
##           4.731693
```

```r
p.val<-pf(var(data1)/var(data2),n1-1,n2-1,lower.tail = F); p.val
```

```
## [1] 0.0006953665
```

Example: One-Tailed F test

```
data1<-c(97,5,8,73,8,88,4,156,80,7,96,3,11,79,5,89,1,119,98,5)
n1<-length(data1); #sd(data1)    #50.00463
data2<-c(25,15,18,43,16,30,2,20,75,13,30,15,21,45,5,89,1,19,50,15)
n2<-length(data2); #sd(data2)   #22.98804
var.test(data1,data2,alternative = "less")
```

```
##
##  F test to compare two variances
##
## data:  data1 and data2
## F = 4.7317, num df = 19, denom df = 19, p-value = 0.9993
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##   0.0000 10.2595
## sample estimates:
## ratio of variances
##           4.731693
```

```
p.val<-pf(var(data2)/var(data1),n1-1,n2-1,lower.tail = T); p.val
```

```
## [1] 0.0006953665
```

Example: Two-Tailed F test

```r
data1<-c(97,5,8,73,8,88,4,156,80,7,96,3,11,79,5,89,1,119,98,5)
n1<-length(data1); #sd(data1)    #50.00463
data2<-c(25,15,18,43,16,30,2,20,75,13,30,15,21,45,5,89,1,19,50,15)
n2<-length(data2); #sd(data2)   #22.98804
var.test(data1,data2)
```

```
##
##  F test to compare two variances
##
## data:  data1 and data2
## F = 4.7317, num df = 19, denom df = 19, p-value = 0.001391
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   1.872862 11.954390
## sample estimates:
## ratio of variances
##           4.731693
```

```r
p.val<-2*pf(var(data2)/var(data1),n1-1,n2-1,lower.tail = T); p.val
```

```
## [1] 0.001390733
```

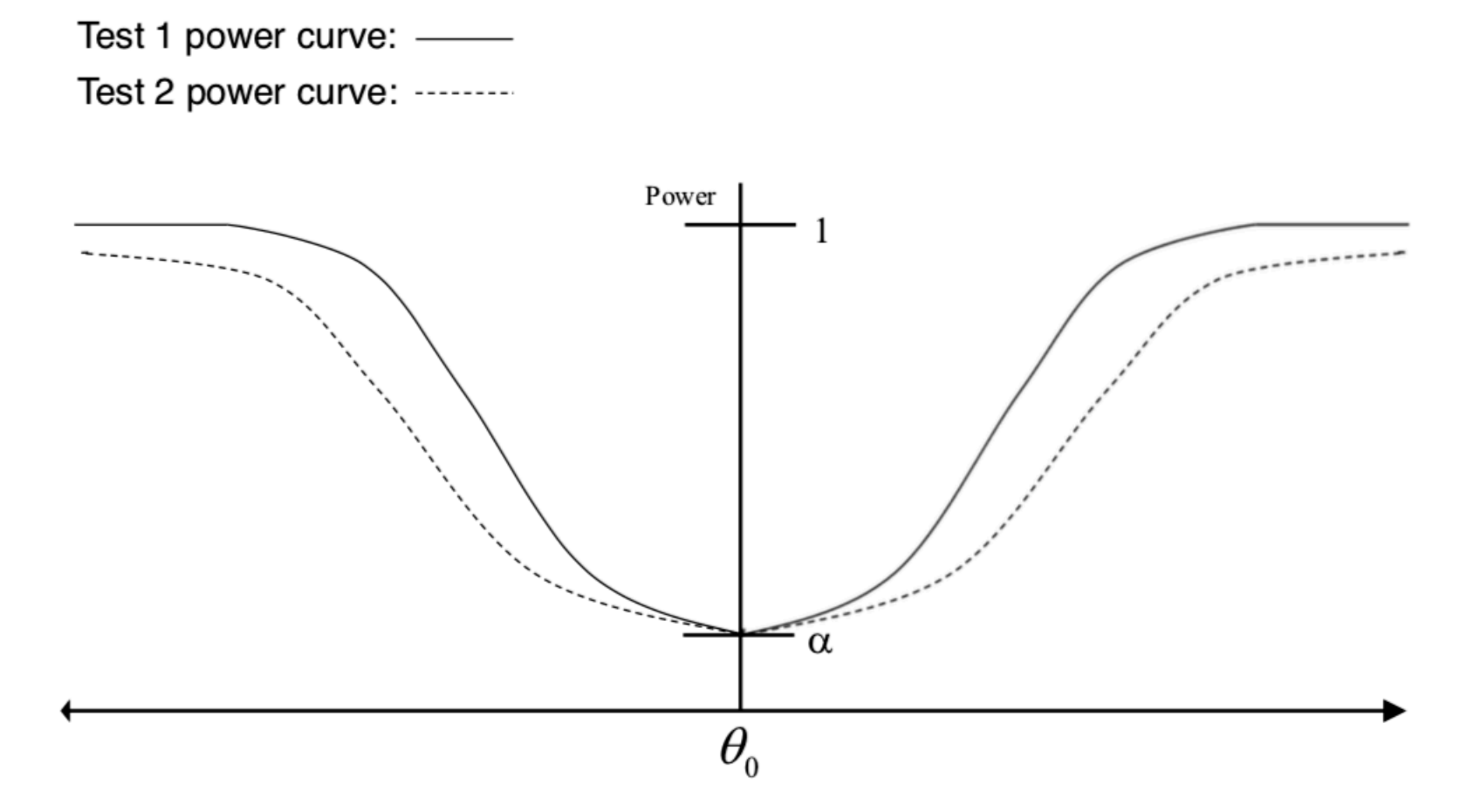
A Note on the Variance Tests

- Unlike the hypothesis tests based on the t distribution, the variance tests are very sensitive to departures from normality.
  – That is, these tests are not robust if the normality assumption is violated
- So, when using these tests, check your data carefully.

How Do We Know Whether All These Tests are Any Good?

- We have learned about how to do hypothesis tests, including evaluating a test's performance in terms of α and β.
- But how do we know the tests we defined are to be preferred?
  – I.e., For any given hypothesis testing scenario, we would want to use the test that achieves the smallest β for a given α.
  – Alternatively, we would want to use the test that achieves the highest power $(1 - \beta)$ for a given α.

Illustrative Power Curves



Simple vs. Composite Hypotheses

- A hypothesis is said to be simple if that hypothesis uniquely identifies the distribution from which the population is taken.
  – Example, when testing the mean of a normally distributed population with σ known, then $H_0 : \mu = \mu_0$ is a simple hypothesis.
- Any hypothesis that is not simple is called a composite hypothesis.
  – Example, for the same test, then $H_a : \mu > \mu_0$ is a composite hypothesis.

More Terminology

- p-value: probability of seeing a test statistic as extreme or more extreme assuming null is true
- Z-test vs t-test.
  – Large sample (z-test) vs. small sample with σ unknown (t-test)
- Simple hypothesis vs. composite hypothesis.
  – Simple hypotheses uniquely specifies the sampling distribution

Neyman-Pearson Lemma

- Neyman-Pearson Lemma: Suppose we want to test a simple null vs. a simple alternative: $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_a$
- Then a test of the form

$$\frac{L(\theta_0)}{L(\theta_a)} < k$$

where k is chosen to achieve a desired level of α, maximizes the power at θ_a.

– That is, the Neyman-Pearon Lemma guarantees a most powerful α-level test

Example

Suppose that we have a random sample of four observations from the density function

$$f(y|\theta) = \begin{cases} \frac{1}{2\theta^3} y^2 e^{-y/\theta} & y > 0, \\ 0 & \text{otherwise} \end{cases}$$

Find the rejection region for the most powerful test of $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_a$, assuming $\theta_a > \theta_0$.

Uniformly Most Powerful α-level Tests

- As we have seen in this and the previous lectures, we usually have composite alternative hypotheses.
- So, ideally, we would like our test to have higher power at every possible alternative.
  – That is, we would like to have a uniformly most powerful test.
- The Neyman-Pearson Lemma only applies to simple vs. simple hypotheses.
  – However, often the RR for a test with a composite alternative only depends on θ_0 (and not θ_a)
  – In these cases we can apply the Neyman-Person Lemma and get a uniformly most powerful test.

Example

Suppose $Y_1, ..., Y_n$ are a random sample from a normal distribution with unknown mean μ and known variance σ^2

- Find the uniformly most powerful α-level test for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$.

Likelihood Ratio Tests

- Method for conducting good hypothesis tests.
  – Similar in spirit to the use of maximum likelihood for constructing good point estimators.
  – Often results in uniformly most powerful tests, or at least tests with reasonably good performance properties
- The Neyman–Pearson lemma is useless if we wish to test a hypothesis about a single parameter θ when the sampled distribution contains other unspecified parameters (nuisance parameter).
- Most tests we have discussed in class actually derived from the likelihood ratio principle.
  – They are called likelihood ratio tests.

Likelihood Ratio Tests

Definition 3

A likelihood ratio test is a test of $H_0 : \Theta \in \Omega_0$ vs $H_a : \Theta \in \Omega_a$, where the rejection region is $\lambda \leq k$ for

$$\lambda = \frac{\max\limits_{\Theta \in \Omega_0} L(\Theta)}{\max\limits_{\Theta \in \Omega} L(\Theta)}$$

- This is a ratio of likelihoods, where the numerator is maximized over the space defined by H_0 and the denominator over the whole space
- It can be shown that $0 \leq \lambda \leq 1$, where smaller values suggest favoring H_a over H_0.

Example

Suppose that $Y_1, ..., Y_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. We want to test $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. Find the appropriate likelihood ratio test.

What we have Just Learned

- Introduction to Testing Hypotheses
  - Steps in Conducting a Hypothesis Test
  - Types of Errors
  - Some terminologies
- Large Sample Tests for the Means and Proportions
  - Z tests
  - Type II Error Probability Calculations
  - Sample Size Calculations
- The p-value approach to hypothesis testing
- Small-sample hypothesis tests (means, proportions)
- Tests for variance(s)
- Neyman-Pearson Lemma and likelihood ratio tests