Statistical Inference and Predictive Analytics

IS 6487

1

Outline

• Discuss Assignment 2

• Estimators and sampling distribution
• Central limit theorem
• Confidence interval
• Hypothesis testing

2

Sample Statistics

- How do we infer about the population statistics such as mean and standard deviation?

- We do so by taking samples and inferring from those samples.
  - Hence we are now moving into the world of inferential statistics.
  - In all cases, the sample size will be denoted *n*.

3

Population Parameters vs. Sample Statistics

- A population parameter is a value that characterizes the population. It is a **fixed** number (assuming the population is static), but it is typically unknown (e.g., population mean).
- A sample statistic is a value that is computed from data derived from a sample. Sample statistics estimate population parameters (e.g., sample mean)

Population

μ: Population mean

σ: Population standard deviation

sampling →

Sample

X = Sample mean

S = Sample standard deviation
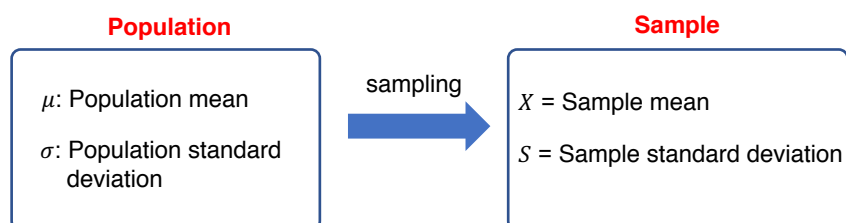
4

Estimator of Population Mean – Sample Mean

- The most widely used sample statistic is sample mean $\bar{X}$.

- Suppose we have sample data $X_1, X_2, \ldots, X_n$ from a population.
- Sample mean is defined as $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

- Sample mean $\bar{X}$ is used to estimate the population mean.

5

Estimator of Population Standard Deviation – Sample Standard Deviation

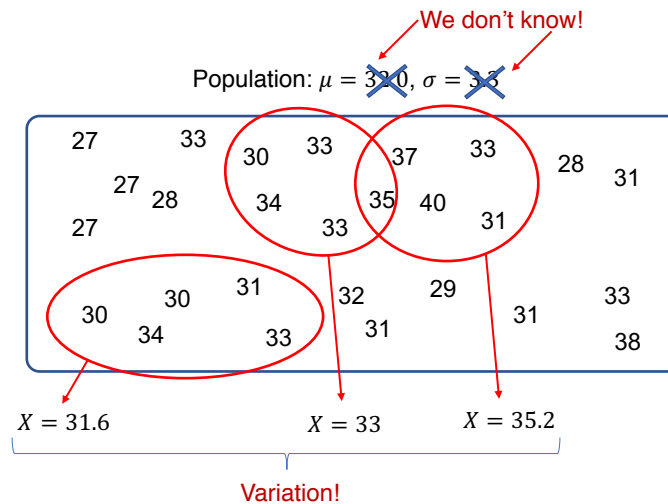- The sample standard deviation, denoted s, is the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Note: s does have the same units as the x_i.

6

Randomness in Sample Mean

- We estimate the population mean by using the sample mean $\bar{X}$

We don't know!

Population: $\mu = 3\!\!\!\times\!\!\!0$, $\sigma = \times$

27   33   30   33   37   33   28   31
  27   28   34   35   40   31
27   33   31
           31   32   29   33
  30   30   31   31
    34   33   38

$X = 31.6$ $X = 33$ $X = 35.2$

Variation!

7

Sample Statistics Vary

- A sample statistic is a **random variable**; hence, its value is not known ahead of time.
  - For example, if I draw one sample and calculate the sample mean, and then draw another sample of the same size and again calculate the sample mean, I will probably obtain a different result.

- If I repeated this procedure many times, and drew a histogram, I would obtain a distribution of values for the sample mean, also known as *sampling distribution*
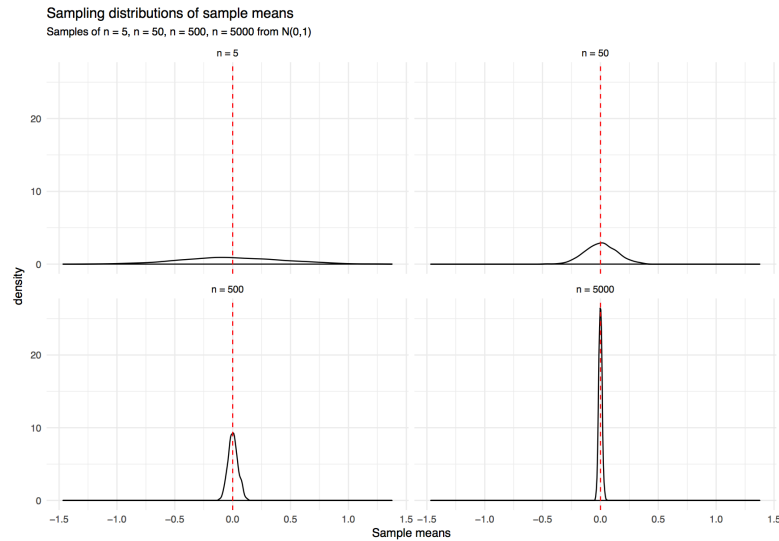
8

Variation in Sampling Distribution (vary sample size)

Sampling distributions of sample means
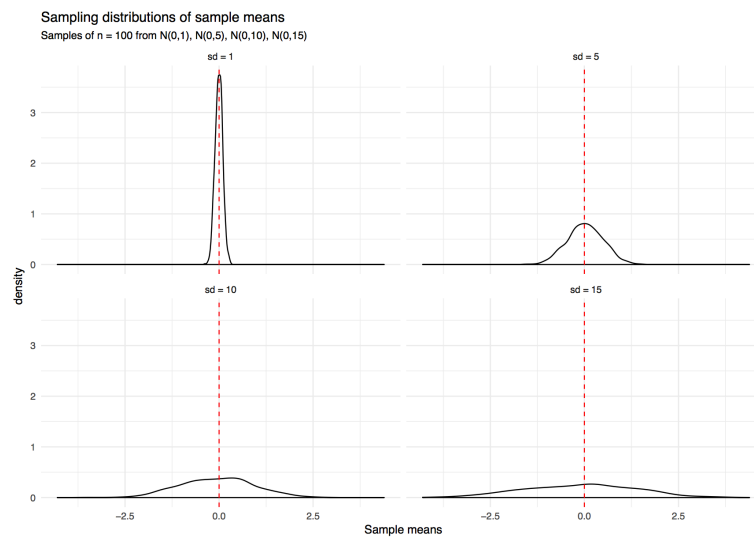Samples of n = 5, n = 50, n = 500, n = 5000 from N(0,1)

Statistics and Predictive Analytics · © B. Kadiyala · 9

9

Variation in Sampling Distribution (vary population variance)

Sampling distributions of sample means
Samples of n = 100 from N(0,1), N(0,5), N(0,10), N(0,15)

Statistics and Predictive Analytics · © B. Kadiyala · 10

10

Sampling Distributions of Sample Mean

1. The sample mean or average $\overline{X}$ is a *random variable* and hence it has a distribution with mean and standard deviation.

2. The mean of the sample averages is equal to the mean of the population of individuals.

3. The spread of the sample averages is always smaller than the spread of the original population of individuals. In fact, the standard deviation (s.d.) of the sample average is

$$\sigma_X = \frac{\sigma}{\sqrt{n}},$$

where σ_X is the s. d. of the sample average, or *standard error*, and σ is the population s. d.

11

Important Characteristics of Sampling Distributions

4. We cannot completely generalize about the shape of the distribution of the sample average. We do know the following, however.
   - If the population from which we are sampling is normal, the sample averages will also have a normal distribution.
   - If we take a sample of size n from any distribution, then as n gets large, the distribution of the sample average approaches a normal distribution. (NOTE that the original distribution does not need to be known.) This is referred to as the *central limit theorem*.

- Rule of thumb
   - The central limit theorem is imprecise about the sample size.
   - A common rule of thumb is to say that if the sample size is at least 30, the central limit theorem applies.

12

Central Limit Theorem

13

Bernoulli distribution

- Bernoulli distribution is a discrete probability distribution of a random variable taking the value 1 ("success") with probability p and the value 0 ("failure") with probability q = 1 − p.

- Bernoulli random variables have **mean p** and **variance p(1 − p).**

14

Sampling from Bernoulli distribution in R

R code for sampling from the Bernoulli distribution:

rbinom(n, size = 1, prob),

where n is the size of the *random sample*, size = 1 indicates a Bernoulli trial, and prob is the probability of success.

15

Sampling from Bernoulli Distribution in R

```
(fair_coin <- rbinom(n = 20, size = 1, prob = .5))
```

[1] 0 0 1 1 1 0 0 1 0 1 0 0 1 1 1 1 0 1 0 1

```
mean(fair_coin)
```

[1] 0.55

```
# mean = p = .5
```

```
var(fair_coin)
```

[1] 0.2605263

```
# variance = p(1 -p) = .5 * .5 = .25
```

16

Sampling from Binomial distribution in R

• R code for sampling from the binomial distribution:

rbinom(n, size, prob),

where n is the size of the random sample, size is the number of Bernoulli trials, and prob is the probability of success.

17

Sampling from Binomial Distribution in R

```
(rep_fair_coin <- rbinom(n = 20, size = 100, prob = .5))

[1] 43 51 59 44 51 49 52 51 53 56 41 47 50 54 43 49 42 41 44 56

mean(rep_fair_coin)

[1] 48.8

# mean = np = 100 * .5 = 50

var(rep_fair_coin)

[1] 29.64211

# variance = np(1 - p) = 100*.5(.5) = 25
```

18

Probability of heads in 3 coin Flips

- In $n = 3$ tosses of a coin there are $2^3 = 8$ possible combinations of heads and tails: HHH, HHT, HTH, THH, HTT, TTH, THT, TTT.
- The probabilities for heads (X = 1, 2, or 3) can then be calculated by simple counting:
  - P(X = 0): TTT or 1/8 (1 instance out of 8 possible combinations)
  - P(X = 1): HTT, TTH, THT or 3/8
  - P(X = 2): HHT, HTH, THH or 3/8
  - P(X = 3): HHH or 1/8

- What do you notice about these probabilities?

19

Probability Mass Function for Binomial(3, .5)

These binomial probabilities look normal!
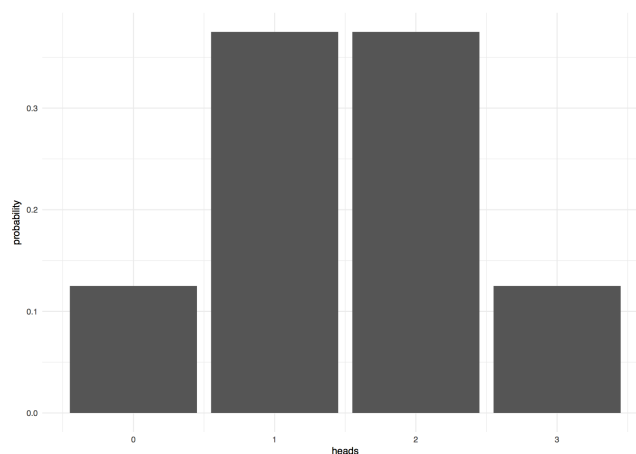
20

Central Limit Theorem

The (sampling) distribution of sample mean approaches a **Normal distribution** as the sample size **increases**.

- The mean of sampling distribution is equal to the population mean
  $\mu_X = \mu$.
- The standard deviation of the sampling distribution is given by
  $\sigma_X = \dfrac{\sigma}{\sqrt{n}}$.
- This holds even when the population distribution is **not** normal

- With the binomial distribution we are summarizing Bernoulli trials by adding up successes, hence the binomial probabilities are normally distributed (roughly at small n, exactly at large n).

Statistics and Predictive Analytics　　　　　　　　　　© B. Kadiyala　　　　　　　　　21

21

Central Limit Theorem – Uniform Distribution

The means are indicated by the dotted red line. These distributions clearly aren't normal.

Statistics and Predictive Analytics　　　　　　　　　　© B. Kadiyala　　　　　　　　　22
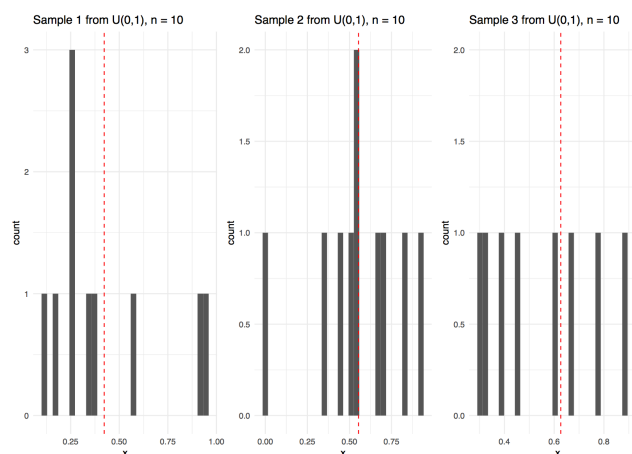
22

Central Limit Theorem – Uniform Distribution

The sampling distribution of means is normal and the mean has converged to the population parameter ($\frac{a-b}{2} = .5$).

Sampling distribution of 1000 means from samples of n = 10
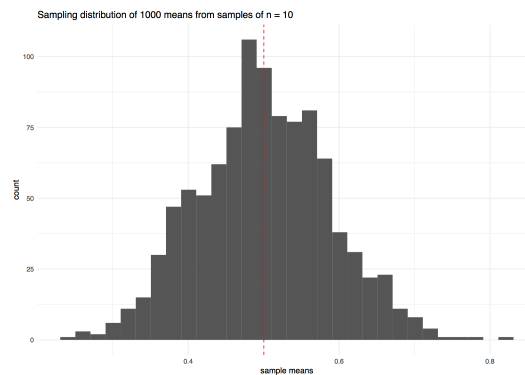
23

Example

• Suppose we are a manufacturer of a product that uses a particular electronic component as a raw material.

• A new supplier comes to us and claims that his component has an average life of 1000 hours.

• We want to verify his claim and so he agrees to allow us to life test 100 of his components.

• Assumption: The time to failure of an electronic component often follows an exponential distribution (and we will assume it does here; as a reminder, the standard deviation of the exponential distribution is equal to its mean.)

24

Example

- **Question**
    - A sample of 100 components is tested, and the sample mean is calculated to be 950 hours.
    - Is the sample result consistent with the supplier's claim? (In other words, if the true mean is 1000 hours, what is the chance of obtaining a sample mean of 950 or smaller?)
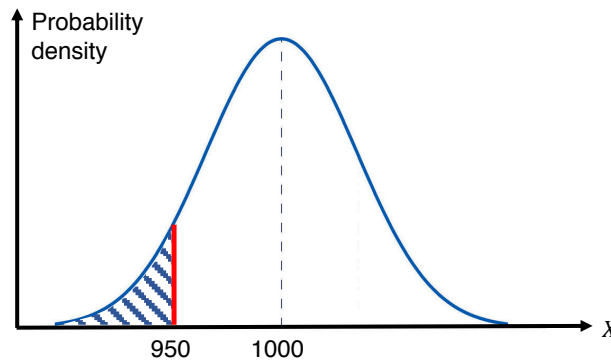
- **Solution**
    - Population mean $\mu = 1000 (= \sigma)$
    - Sample mean $\bar{X} = 950$
    - Standard deviation of sample mean $\sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{1000}{\sqrt{100}} = 100$
    - Distribution of $\bar{X}$ **approximately** normal using CLT
    - We need to find out the probability of $\bar{X} \leq 950$

25

Example

- Answer
```
> pnorm(950,1000,100)
[1] 0.3085375
```
- That is, the probability that $\bar{X} \leq 950$ is 31% *if the true mean is indeed 1000*

26

13

A Note on Different Standard Deviations

- The population standard deviation, $\boldsymbol{\sigma}$
  - The spread of all individuals in the population. Typically unknown.

- The sample standard deviation, $\boldsymbol{s}$
  - Estimator of σ. Obtained from the spread of individuals in the sample.

- The standard error or standard deviation of the sample mean, $\boldsymbol{\sigma_X}$
  - The spread of the sample mean, $\boldsymbol{\sigma_X = \sigma/\sqrt{n}}$. Approaches zero as the sample size n increases. Also typically unknown because $\boldsymbol{\sigma}$ is unknown.

- The estimated standard error, $\boldsymbol{s_X}$
  - An estimate of the spread in sample means, $\boldsymbol{s_X = s/\sqrt{n}}$.

27

Steps in Statistical Inference

1. Quantity of interest, e.g., population mean (unknown but **not** a r.v.)
2. Sample from population
3. Find an *estimator*, e.g., sample mean is the best estimator for population mean
4. **Determine the distribution of the estimator (*sampling distribution*)**
   i. **If the population distribution is normal we have exact sampling distributions**
   ii. **If the population distribution is not normal then we need large enough sample size to approximate using a Normal distribution (CLT)**
5. Make inference using the *estimate* and the sampling distribution

28

Assumptions

- We will assume the population is infinitely large
  - Sample size is finite

- The population distribution is Normal
  - If not, we can invoke the CLT to do asymptotic analysis
  - Normal distribution gives us **exact** sampling distributions

- We will assume that the population standard deviation is unknown
  - This represents most real-world problems

29

Confidence Intervals

30

Point Estimate

- If we estimate a parameter with a single number, it is called a *point estimate*. The statistic that is used to find that single number is called a *point estimator*.

- The sample mean, $\bar{X}$, is a point estimator of the population parameter (mean), μ.

- Estimators are random variables, parameters are not.

31

Point Estimate – Precise but not Accurate

- While a point estimate of a parameter is precise, it is unlikely that it will be exactly right.

- Therefore we can also use interval estimates which are less precise, but have a better chance of being correct.

- These interval estimates are referred to as *confidence intervals*.

32

Confidence Intervals

• Nearly all confidence intervals have the same following form.

point estimator ± (*multiplier*) x (*standard error of the point estimator*)

• The multiplier depends on the distribution of the point estimator. We use the multiplier to obtain the confidence level that we want.

33

Computing the Interval (continued)

• Suppose you have a sample of observations $X_1, X_2, ..., X_n$ where the population distribution is normal with mean μ and standard deviation σ (both **unknown**), then the normalized sample mean is t-distribution

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \text{ has t-distribution.}$$

• This is **not an approximation** rather an exact result.

34

17

Sampling Distribution of Sample Mean – Student t-distribution
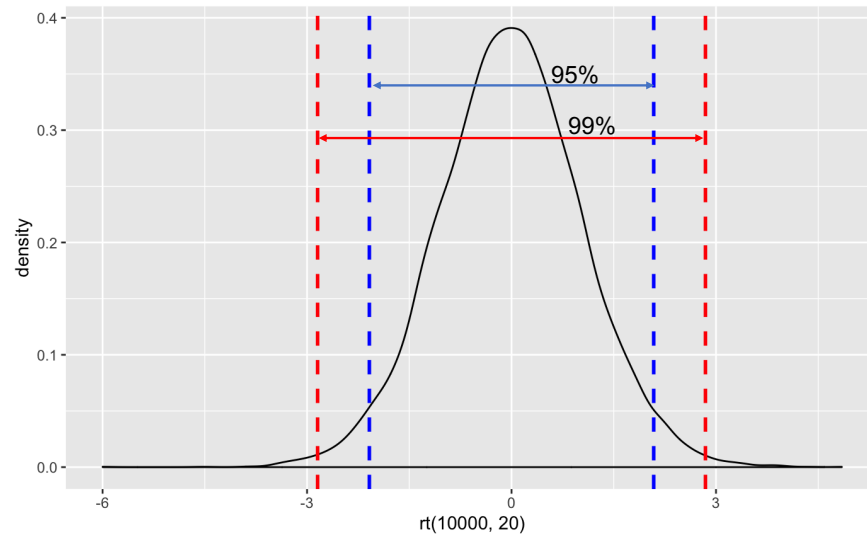
35

## Sampling Distribution to Inference

Say you observe a t-statistic of 1.6 with a sample size of 20. What can you infer?
Suppose you determine t_{CL} corresponding to 95%

$$\text{Prob}\left(\frac{X-\mu}{\frac{s}{\sqrt{n}}} \in [-t_{CL}, t_{CL}]\right) = 95\%$$

$$\Leftrightarrow \text{Prob}(\mu - t_{CL}\frac{s}{\sqrt{n}} \leq \bar{X} \leq \mu + t_{CL}\frac{s}{\sqrt{n}}) = 95\%$$

$$\Leftrightarrow \text{Prob}(\bar{X} - t_{CL}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{CL}\frac{s}{\sqrt{n}}) = 95\%$$

- That is, the **random** interval $[\bar{X} - t_{CL}\frac{s}{\sqrt{n}}, \bar{X} + t_{CL}\frac{s}{\sqrt{n}}]$ contains **true** population mean **95%** of the time!
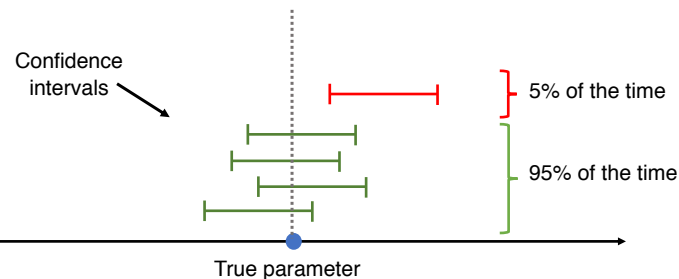 - Also referred to as 95% *confidence interval* for population mean
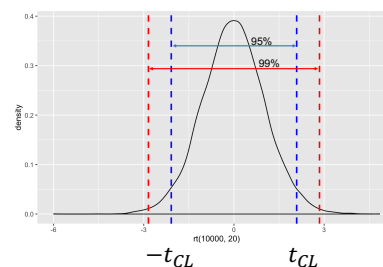
36

Illustration

- Suppose the blue dot is the true parameter value.
- 95% confidence interval means that the confidence interval will contain the true parameter value 95% of the time.

37

Computing the Interval

- From the definition of confidence level, we want the probability that our interval covers the mean to be *CL*.

- Let's call the lower limit of the interval L and the upper limit R. Then we want

$$L = \bar{X} - t_{CL}\frac{s}{\sqrt{n}} \quad \text{and} \quad R = \bar{X} + t_{CL}\frac{s}{\sqrt{n}}$$

$$P(L \leq \mu \leq R) = CL.$$

- Note that t_{CL} changes depending on sample size n and confidence level needed.
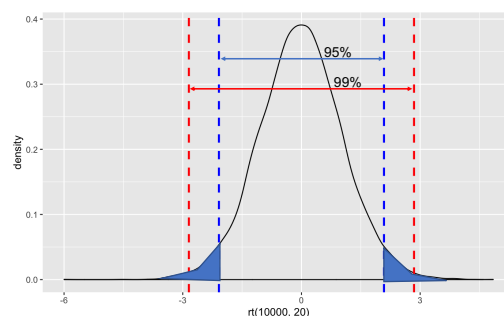
38

t Distribution in R

df = n-1 is called the degrees of freedom for a sample of size n.

- dt(x, df): probability density at value x

- pt(q, df): probability that the r.v. can take less than or equal to x

- qt(p, df): pth-quantile

- rt(n, df): randomly samples from t-distribution

39

Computing Confidence Limits

- t_{CL} corresponding to 95% confidence level with a sample size 20


```
qt(.025,20)
qt(.975,20)

` ` `

[1] -2.085963
[1] 2.085963
```

both areas in blue =.025 thus the total area between the blue lines is 1-.05 = .95

- When sample size is more than 30 you use normal instead of distribution.

40

Null Hypothesis Statistical Testing (NHST)

Preview of The Procedure

• To structure the hypothesis testing procedure, we can follow a few steps.

1. Formulate the null and alternative hypotheses and determine if the test is 1 or 2 sided.
2. Set the sample size, collect the data and compute the appropriate test statistic.
3. Determine the level of error that can be tolerated (α).
4. Find the p-value.
5. Make the conclusion using the following rule: If the p-value is less than α, reject the null hypothesis.

• Note that testing a statistical hypothesis is essentially equivalent to following the scientific method.

Step 1: Establishing the Hypotheses

- A statistical hypothesis is an assumption or statement concerning one or more populations.

- Most often we make and test hypotheses about the parameters of a particular distribution.

- We will always formulate two hypotheses.
  - The first is called the null hypothesis, and is denoted H_0.
  - The second is called the alternative hypothesis, and is denoted H_a.

43

Step 1: Establishing the Hypotheses

Guidelines for the Hypotheses

- When forming the hypotheses, keep in mind that we can only prove something by rejecting the null hypothesis.
- As a result, we usually form the null hypothesis to try to reject it.
- *In other words, whatever we want to substantiate from the data typically goes into the alternative hypothesis.*
- Another guideline is that the null hypothesis is typically based on the status quo, or historical value.

44

Step 1: Establishing the Hypotheses

One- vs. Two-Sided Tests
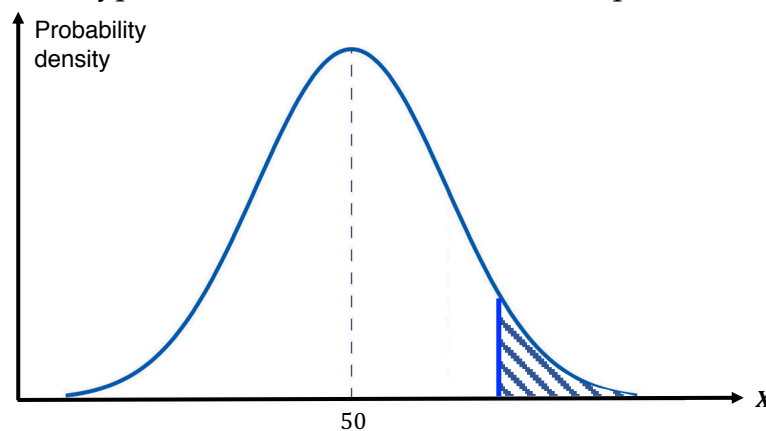
• The form of the *alternative* hypothesis is important in what we develop later.

• If the alternative hypothesis has the form "not equal to" ($\neq$, usually indicated in the verbiage of the problem as the parameter has "changed"), then the test is said to be a 2-tailed (or 2-sided) test.

• If the alternative hypothesis is less than (<, indicated by verbiage suggesting that the parameter has decreased) or greater than (>, indicated by verbiage suggesting that the parameter has increased), the test is 1-tailed (or 1-sided).

45

Step 1: Establishing the Hypotheses

Picturing the "Tails"

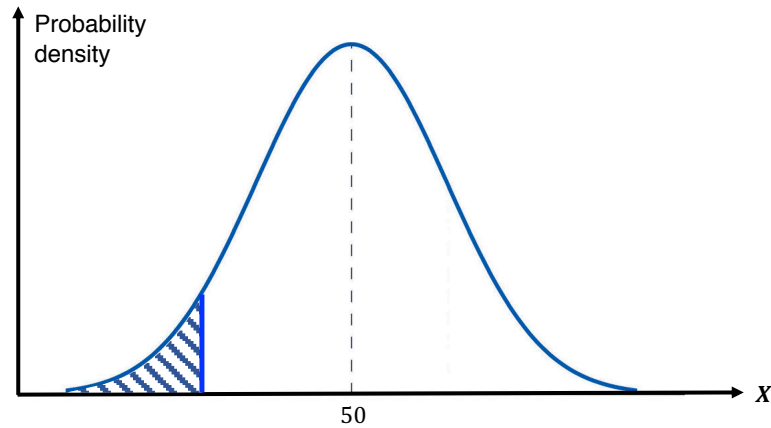• The alternative hypothesis H_a includes ">". For example, **H_a: $\mu > 50$**

46

Step 1: Establishing the Hypotheses

Picturing the "Tails"

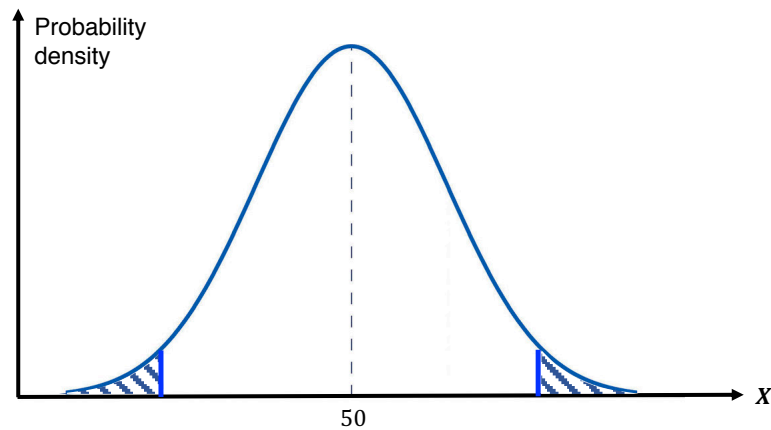• The alternative hypothesis H_a includes "<". For example, **H_a: $\mu < 50$**

47

Step 1: Establishing the Hypotheses

Picturing the "Tails"

• The alternative hypothesis H_a includes "≠". For example, **H_a: $\mu \neq 50$**

48

24

Step 1: Establishing the Hypotheses

Example 1

- A health care provider claims that the mean overpayment on claims from an insurance company was 0.

- The insurance company is skeptical, believing that they were overcharged. They want to use sample data to prove their belief.

- A sample (audit) of 36 claims was collected, and a sample mean overpayment of $4 was calculated with a standard deviation of $12.

- Suppose that we can tolerate a chance of 5% of rejecting the claim when it is true. Is there evidence in the data to conclude that the insurance company was overcharged?

49

Step 1: Establishing the Hypotheses

Example 1

- Which tail(s) are we interested in?

- H_0: Overpayment = 0

- H_a: Overpayment > 0

50

Step 2: Collecting Data and Computing Test Statistic

Determining the Appropriate Test Statistic

- Once we collect the sample data, we compute the appropriate *test statistic*.

- For most of the cases (including all of those that we will talk about), the test statistic tells us the distance (measured in standard errors) between what we observe in the data and what is hypothesized.

51

Step 2: Collecting Data and Computing Test Statistic

The t-Statistic

- If we are interested in the **mean** (and we do not know the population standard deviation), the *test statistic* is

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

  where μ is the hypothesized mean value and the degrees of freedom is $n - 1$

For example, we might test if μ is equal to zero. Null hypothesis would be $\mu = 0$

52

Step 2: Collecting Data and Computing Test Statistic

Example

- Recall in our example that 36 claims were audited.
- The sample mean and standard deviation were calculated to be $4 and $12 respectively.
  - $T = \frac{X - \mu_0}{s/\sqrt{n}} = \frac{4-0}{12/\sqrt{36}} = 2$

- Interpretation:
  - The difference between the sample mean and the hypothesized value is 2 standard errors.

53

Step 2: Collecting Data and Computing Test Statistic

The two-sample t-Statistic

If we are interested in comparing **two** population **means** (and we do not know the population standard deviation), the *test statistic* is

$$T = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $(\mu_1 - \mu_2)$ is the true difference in means and the degrees of freedom is given by

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

For example, we might test if $(\mu_1 - \mu_2) > 0$, i.e., the mean in population 1 is higher than that of population 2

54

Step 3. Setting the Level of Significance

How far is far?

- Sample statistics will exhibit variation ("stuff varies!"), and we need to know how much variation we can tolerate.

- Intuitively, for our example, we would reject the null hypothesis if the sample statistic $\bar{X}$ is sufficiently large.

- But how large? Is $\bar{X}$=2 large enough?   What about 3? The theory behind hypothesis testing gives us the tools to answer these questions.

55

Step 3. Setting the Level of Significance

- To determine if a sample statistic is sufficiently large or small to reject the hypothesis, we need to determine how likely it is to obtain the sample statistic value that we found in Step 2 *if the null hypothesis is true*.

- If it is not very likely, we won't believe the null hypothesis.

- If it is likely, we will not have enough evidence to reject the null hypothesis. In other words, what we observe in the data will be consistent with what we have hypothesized.

56

Step 3. Setting the Level of Significance

- We will use the distribution of the sample statistic in question to determine the desired probability.
- To decide if the probability is too small, we need to define the types of errors that can be made when testing a hypothesis.

	True (but Unknown) State	
Conclusion	H_0 **is true**	H_0 **is false**
Reject H0	**Type I error** (α)	
Do not reject H0		**Type II error** (β)

- The probability of a Type I error is denoted α, and that of a Type II error is denoted β.
- 1 - β is referred to as the power of the test.

57

Step 3. Setting the Level of Significance

The level of significance
- To test the hypothesis, we control α by setting it to be a small number.
- The most typical number chosen is .05. Other numbers that are common are .01 and .10.
- While it is the probability of a Type I error, the most common term for α is the *level of significance*.

For our example
- The problem statement for our example said, "Suppose that we can tolerate a chance of 5% of rejecting the claim when it is true."
- In other words, we want the level of significance of our test to be .05.
- How do we represent the significance level pictorially?

58

29

Step 3. Setting the Level of Significance

Pictorial representation for our example
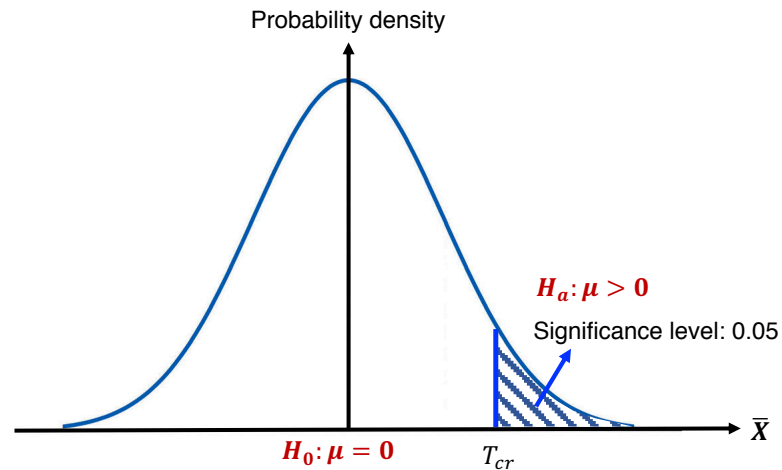
59

Step 4: Finding the p-value

- The fourth step is to determine if the sample value we obtained is consistent with the null hypothesis.
  - If it is consistent, we will not reject the null hypothesis and conclude that the sample statistic could have come from a distribution with the hypothesized parameter.
  - If it is not, we will reject the null hypothesis in favor of the alternative.

What is a p-value anyway?

- Probability that the difference we observe between the hypothesized value and the sample statistic is due to chance
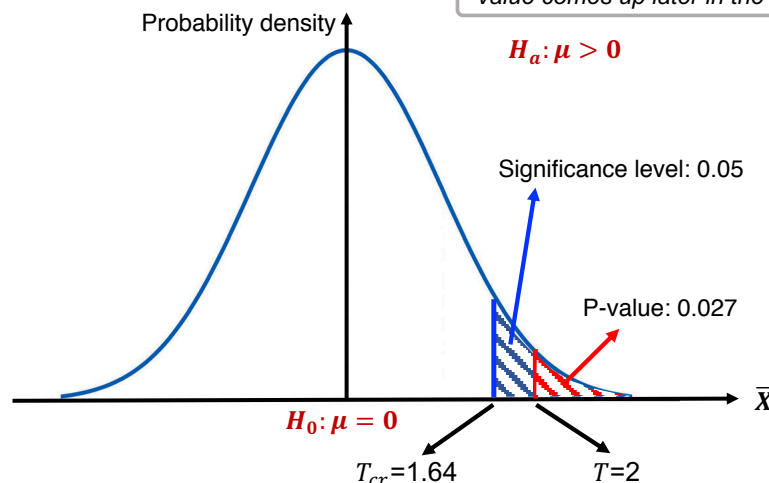
60

Step 4: Finding the p-value

Recall Example 1

- To illustrate, we will use the numbers from our example.
- Recall that the hypothesized mean was 0 and the sample mean was 4.
    - The T-statistic was **2**.
- The alternative hypothesis was that the mean was greater than 0.

61

Step 4: Finding the p-value

Recall Example 1

How to actually calculate the p-value comes up later in the slides

$H_a: \mu > 0$

Probability density

Significance level: 0.05

P-value: 0.027

$H_0: \mu = 0$

$T_{cr}=1.64$ $T=2$

$\bar{X}$

62

Step 4: Finding the p-value

Computing p-value

- There are 3 cases to consider when finding the p-value. The cases are all based on the form of the alternative hypothesis.

- **Case 1**: If the alternative hypothesis is of the form "<," then we calculate the left tail probability.
- **Case 2**: If the alternative hypothesis is of the form ">," then we calculate the right tail probability.
- **Case 3**: If the test is two-sided (of the form ≠), we find the one tail area and multiply it by 2.
  - If the test statistic value is positive, we find the right tail area and multiply it by 2.
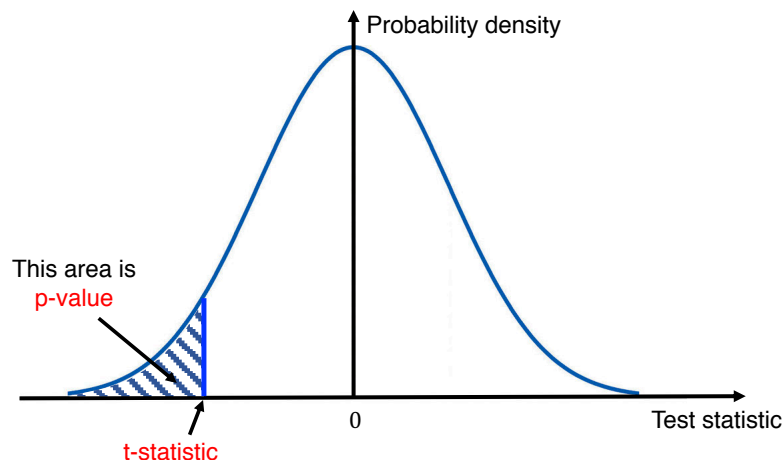 - If the test statistic value is negative, we find the left tail area and multiply it by 2.

63

Step 4: Finding the p-value

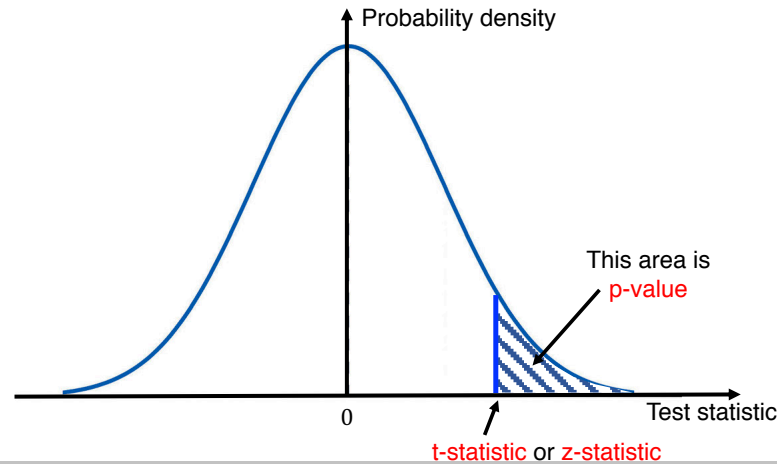- **Case 1:** If the alternative hypothesis is of the form "<," then we calculate the left tail probability. For example, $\mathbf{H_a: \mu < 50}$

64

Step 4: Finding the p-value

- **Case 2**: If the alternative hypothesis is of the form ">," then we calculate the right tail probability. For example, H_a: $\mu > 50$

Probability density

This area is p-value

0

Test statistic

t-statistic or z-statistic
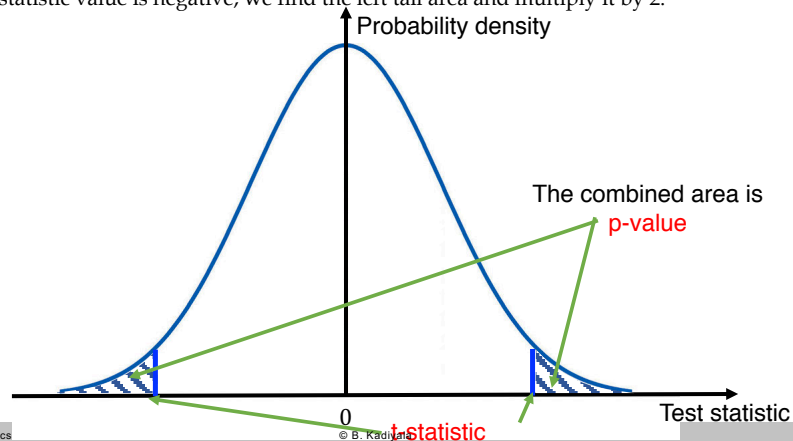
65

Step 4: Finding the p-value

- **Case 3**: If the test is two-sided (of the form ≠), we find the one tail area and multiply it by 2. For example, H_a: $\mu \neq 50$
  - If the test statistic value is positive, we find the right tail area and multiply it by 2.
  - If the test statistic value is negative, we find the left tail area and multiply it by 2.

Probability density

The combined area is p-value

0

t-statistic

Test statistic

66

Step 4: Finding the p-value

Computing p-value: Just find the tail areas

- So finding the p-value amounts to finding tail areas.

- If we are testing a mean, then we use a **t-statistic** so we need to find the tail(s) of the t-distribution.

- If we are testing a proportion, we use a **z-statistic** so we need to find the tail(s) of the standard normal distribution.

67

Step 4: Finding the p-value

- We can compute p-values using

>pt(statistic value, df)

- If degree of freedom is very high (>30) we can approximate t-distribution using normal

>pnorm(statistic value)
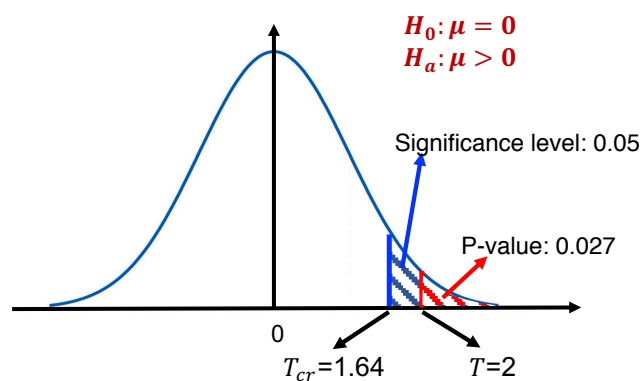
68

Step 4: Finding the p-value

Recall our Example 1

- Recall that in our example the alternative hypothesis was μ>0.
- Also recall that the sample size was 36 and the sample mean was $4, from which we calculated T=2.
- With that information, what is the p-value for the example?

69

Step 4: Finding the p-value

- Recall our Example

$$H_0: \mu = 0$$
$$H_a: \mu > 0$$

```
> 1-pt(2,35)
[1] 0.02665383
> 1-pnorm(2)
[1] 0.02275013
```

Significance level: 0.05

P-value: 0.027

0

$T_{cr} = 1.64$ $T = 2$

- Interpretation: If H_0 (null hypothesis) is true, the probability of getting a sample mean of $4 or higher is 0.027.

70

Step 5: Making Conclusions

- To conclude the hypothesis testing procedure, we combine the level of significance from Step 3 with the p-value from Step 4.

- The rule we will use is always the same, no matter what the form of the alternative hypothesis.

Reject H_0 if the p-value is less than α.

- It is useful to make two conclusions at the end of the process
  1. The first conclusion is to reject or not reject the null hypothesis.
  2. The second conclusion is an interpretation of what it means to reject (or not) in the context of the problem.