DATA SCIENCE WITH R

DATA SCIENCE WITH R

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

Population Distribution Not Normal

Population Distribution Not Normal

<u>Every hypothesis test may not use a normal distribution</u>

Population Distribution Not Normal

<u>Every hypothesis test may not use a normal distribution</u>

Example:

A manufacturer claims 2 out of 5 people prefer their washing powder over any other brand. A random sample of 25 people results in 4 people preferring this brand. Is the manufacturer's claim justified? Test at 95% confidence

Population Distribution Not Normal

Every hypothesis test may not use a normal distribution

Example:

A manufacturer claims 2 out of 5 people prefer their washing powder over any other brand. A random sample of 25 people results in 4 people preferring this brand. Is the manufacturer's claim justified? Test at 95% confidence

What would be null hypothesis here?

Population Distribution Not Normal

<u>Every hypothesis test may not use a normal distribution</u>

Example:

A manufacturer claims 2 out of 5 people prefer their washing powder over any other brand. A random sample of 25 people results in 4 people preferring this brand. Is the manufacturer's claim justified? Test at 95% confidence

What would be null hypothesis here?

Ho: Brand preference is 40% (2/5)

Population Distribution Not Normal

<u>Every hypothesis test may not use a normal distribution</u>

Example:

A manufacturer claims 2 out of 5 people prefer their washing powder over any other brand. A random sample of 25 people results in 4 people preferring this brand. Is the manufacturer's claim justified? Test at 95% confidence

What would be null hypothesis here?

Ho: Brand preference is 40% (2/5)
H1: Brand preference less than 40%

Population Distribution Not Normal

Every hypothesis test may not use a normal distribution

Example:

A manufacturer claims 2 out of 5 people prefer their washing powder over any other brand. A random sample of 25 people results in 4 people preferring this brand. Is the manufacturer’s claim justified? Test at 95% confidence

What would be null hypothesis here?

Ho: Brand preference is 40% (2/5)
H1: Brand preference less than 40%

Sig Level: 5%

Population Distribution Not Normal

<u>Every hypothesis test may not use a normal distribution</u>

Example:

A manufacturer claims 2 out of 5 people prefer their washing powder over any other brand. A random sample of 25 people results in 4 people preferring this brand. Is the manufacturer's claim justified? Test at 95% confidence

What would be null hypothesis here?

Ho: Brand preference is 40% (2/5)
H1: Brand preference less than 40%

Sig Level: 5%

Test Distribution?

Population Distribution Not Normal

The outcome in the population is Binomially Distributed (Prefer / Do Not Prefer)

Binomial Distribution:

P (Seeing a 16% or less preference rate, when expecting 40%)

Population Distribution Not Normal

The outcome in the population is Binomially Distributed (Prefer / Do Not Prefer)

Binomial Distribution:

P (Seeing a 16% or less preference rate, when expecting 40%)

= Binom.dist(4,25,0.4,true) = 0.009

Population Distribution Not Normal

The outcome in the population is Binomially Distributed (Prefer / Do Not Prefer)

Binomial Distribution:

P (Seeing a 16% or less preference rate, when expecting 40%)

= Binom.dist(4,25,0.4,true) = 0.009

Since p-value < Sig Level (5% = 0.05), we REJECT the null hypothesis

Population Distribution Not Normal

The outcome in the population is Binomially Distributed (Prefer / Do Not Prefer)

Binomial Distribution:

P (Seeing a 16% or less preference rate, when expecting 40%)

= Binom.dist(4,25,0.4,true) = 0.009

Since p-value < Sig Level (5% = 0.05), we REJECT the null hypothesis

Conclusion:

Manufacturer’s claim is NOT justified, and brand preference is actually less than 40%, at a 95% level of confidence

Population Distribution Not Normal

We could alternatively use a normal distribution -

Population Distribution Not Normal

We could alternatively use a normal distribution -

- If we have a binomially distributed random variable:

  approx. mean = n*p
  approx. std deviation = (npq)^0.5

Population Distribution Not Normal

We could alternatively use a normal distribution -

- If we have a binomially distributed random variable:

approx. mean = n*p
approx. std deviation = (npq)^0.5

Here **mean** = 25*0.4 = 10
stdev = (0.4*25 *0.6)^0.5 = 2.44

Population Distribution Not Normal

We could alternatively use a normal distribution -

- If we have a binomially distributed random variable:

approx. mean = n*p
approx. std deviation = (npq)^0.5

Here **mean** = 25*0.4 = 10
stdev = (0.4*25 *0.6)^0.5 = 2.44

- Normal distribution formula -

P = norm.dist(4,10,2.44,true) = 0.006: Conclusion- Reject Null

Population Distribution Not Normal

We could alternatively use a normal distribution -

- If we have a binomially distributed random variable:

approx. mean = n*p
approx. std deviation = (npq)^0.5

Here **mean** = 25*0.4 = 10
stdev = (0.4*25 *0.6)^0.5 = 2.44

- Normal distribution formula -

P = norm.dist(4,10,2.44,true) = 0.006: Conclusion- Reject Null

* Not really appropriate to use a normal distribution because sample size < 30

DATA SCIENCE WITH R

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

Sample Sizes are Low

Hypothesis Testing T-Tests

Hypothesis Testing T-Tests

<u>Example</u>:

We test if college students sleep a lot less than the general population - average sleep hours for the population is 8 hours.

Hypothesis Testing T-Tests

Example:

We test if college students sleep a lot less than the general population - average sleep hours for the population is 8 hours.

Taking a random sample of 10 college students, we get this data.

Student	Sleep Hrs
1	7
2	6.8
3	6
4	7
5	5.5
6	6.6
7	5.5
8	7.5
9	9
10	5.5
Avg	**6.64**

Hypothesis Testing T-Tests

<u>Example</u>:

We test if college students sleep a lot less than the general population - average sleep hours for the population is 8 hours.

Taking a random sample of 10 college students, we get this data.

Should we conclude that students sleep less than the general population?

Student	Sleep Hrs
1	7
2	6.8
3	6
4	7
5	5.5
6	6.6
7	5.5
8	7.5
9	9
10	5.5
Avg	**6.64**

Hypothesis Testing T-Tests

- In order to compute probability of an observed outcome when sample size < 30, the sample means follow what is called a **T-Distribution**

Hypothesis Testing T-Tests

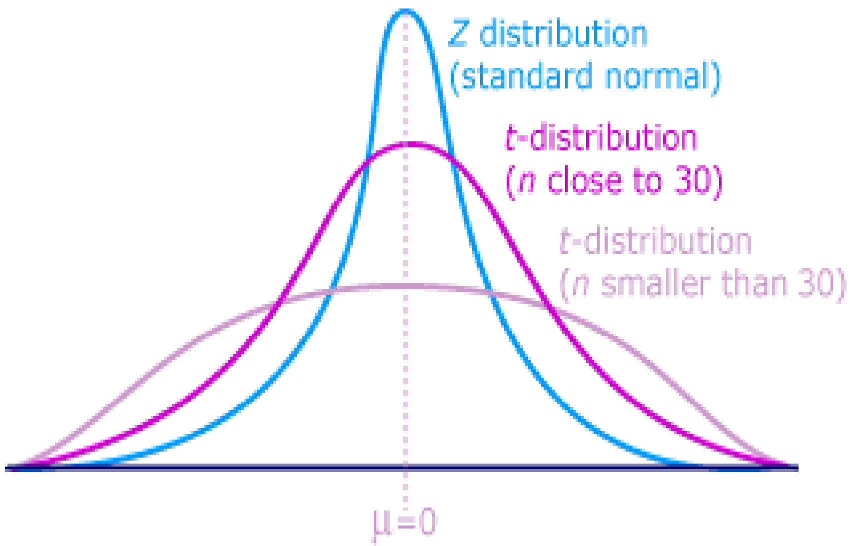
- In order to compute probability of an observed outcome when sample size < 30, the sample means follow what is called a **T-Distribution**

- Do not use the Central Limit Theorem normal approximation because it holds good for sample sizes of at least 30

Hypothesis Testing T-Tests

- In order to compute probability of an observed outcome when sample size < 30, the sample means follow what is called a **T-Distribution**

- Do not use the Central Limit Theorem normal approximation because it holds good for sample sizes of at least 30



Hypothesis Testing T-Tests

For a random sample of size *n* (less than 30) drawn from a population with mean μ and standard deviation sigma:

Hypothesis Testing T-Tests

For a random sample of size *n* (less than 30) drawn from a population with mean μ and standard deviation sigma:

1. The distribution of sample means has a *t* distribution with ***n*-1** degrees of freedom

Hypothesis Testing T-Tests

For a random sample of size *n* (less than 30) drawn from a population with mean μ and standard deviation sigma:

1. The distribution of sample means has a *t* distribution with ***n*-1** degrees of freedom

2. As sample size increases and approaches 30, the t-dist approximates a normal distribution

Hypothesis Testing T-Tests

For a random sample of size *n* (less than 30) drawn from a population with mean μ and standard deviation sigma:

1. The distribution of sample means has a *t* distribution with ***n*-1** degrees of freedom

2. As sample size increases and approaches 30, the t-dist approximates a normal distribution

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Hypothesis Testing

Null Hypothesis

Alternate

Significance Level

Test Statistic

Critical Distance

Hypothesis Testing

Null Hypothesis:
Ho: Students sleep same as population

Hypothesis Testing

Null Hypothesis:
Ho: Students sleep same as population

Alternate:
H1:Students sleep < population

Hypothesis Testing

Null Hypothesis:
Ho: Students sleep same as population

Alternate:
H1:Students sleep < population

Significance Level: 5%

Hypothesis Testing

Null Hypothesis:
Ho: Students sleep same as population

Alternate:
H1:Students sleep < population

Significance Level: 5%

Test Statistic:
(6.44 – 8)/(1.1/(10^0.5)) = -3.90

Hypothesis Testing

Null Hypothesis:
Ho: Students sleep same as population

Alternate:
H1:Students sleep < population

Significance Level: 5%

Test Statistic:
(6.44 – 8)/(1.1/(10^0.5)) = -3.90

Critical Distance:
5%, 9 df, one tail = 1.833

Hypothesis Testing

Null Hypothesis:
Ho: Students sleep same as population

Alternate:
H1:Students sleep < population

Significance Level: 5%

Test Statistic:
(6.44 – 8)/(1.1/(10^0.5)) = -3.90

Critical Distance:
5%, 9 df, one tail = 1.833

What happens when Test Statistic is negative?

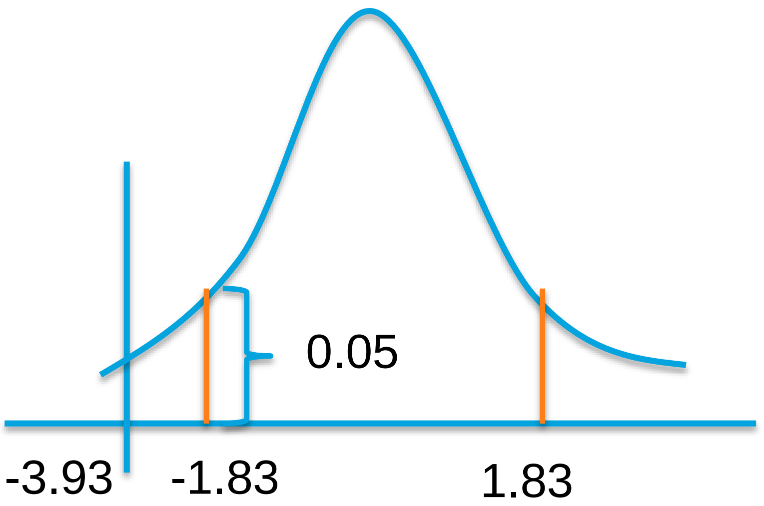
alpha one-tailed	.05	.025	.01	.005
alpha two-tailed	.10	.05	.02	.01
df				
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.743	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.869	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
inf	1.645	1.96	2.326	2.576

Hypothesis Testing

Critical Values calculated based on cut-off probabilities of outcomes to the right of the mean

If test statistic is negative, it simply implies that sample mean is < pop mean

Critical Value
=1.83 to right of mean=-1.83 to left of mean



Distribution tables usually show cumulative probabilities from infinity to Z

Hypothesis Testing

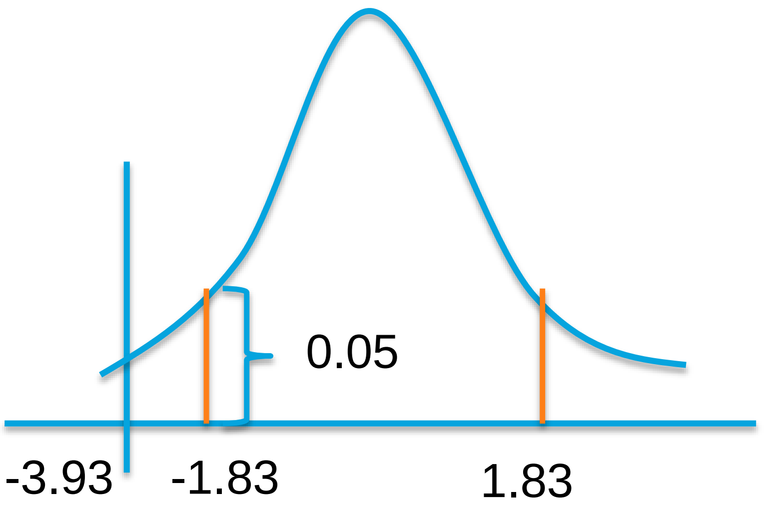
Critical Values calculated based on cut-off probabilities of outcomes to the right of the mean

If test statistic is negative, it simply implies that sample mean is < pop mean

Critical Value
=1.83 to right of mean=-1.83 to left of mean

If test statistic is farther away from mean than critical value, reject null

Conclusion - Students sleep less than general population



Distribution tables usually show cumulative probabilities from infinity to Z

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Step 1:

Step 2:

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Step 1: Calculate the T-Distance: $t = \dfrac{\overline{x} - \mu}{s / \sqrt{n}}$

Step 2:

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Step 1: Calculate the T-Distance: $t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$

Step 2: Use the T-Distance value in Excel with the following formula:
T.DIST(T-Distance, Degrees of Freedom, TRUE)

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Step 1: Calculate the T-Distance: $t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$

Step 2: Use the T-Distance value in Excel with the following formula:
T.DIST(T-Distance, Degrees of Freedom, TRUE)

Degrees of Freedom = n-1

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Step 1: Calculate the T-Distance: $t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$

Step 2: Use the T-Distance value in Excel with the following formula:
T.DIST(T-Distance, Degrees of Freedom, TRUE)

Degrees of Freedom = n-1

In our example: p value of outcomes more extreme than observed =
T.DIST(-3.90, 9, TRUE) = 0.00181

Hypothesis Testing T-Tests

We can directly calculate p-value using the T-Distribution pdf in Excel:

Step 1: Calculate the T-Distance: $t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$

Step 2: Use the T-Distance value in Excel with the following formula:
T.DIST(T-Distance, Degrees of Freedom, TRUE)

Degrees of Freedom = n-1

In our example: p value of outcomes more extreme than observed =
T.DIST(-3.90, 9, TRUE) = 0.00181.

Reject the null hypothesis and conclude that students sleep < general population

Coming Up

Types of Hypothesis Tests:

Population Std Deviation not known

THANK YOU