Statistics-Concepts Explanation[2]

Friday, March 15, 2019

9:46 PM

[1] z- Score

[2] t-score

[3] Central Limit Theorem

[4] Hypothesis Testing

[5] ANOVA

[6] Correlation Coefficient

**Calculating Random (Chance) Probability**

**[1] z-value (score)**

- is a measure of identifying the standard deviation away or below from the population mean. The z-score on the normal distribution curve, ranges from -3 standard deviations up to +3 standard deviations. e.g. z value = 1.2 is interpreted as observed value is +1.2 standard deviations away from the mean.

* z-score has been used to calculate the probability of occurring the score in the normal distribution.
* z-score is a way to compare results from a test to "normal" population. i.e. experiment of survey will have thousands of possible results and many seems to be meaningless.

Machine generated alternative text:
Here X is the point on the curve, 
(T 
Mean 
(J Standard Deviation 
p is mean of the population and o is standard deviation of population. 
Standard Normal Distribution 
Standard Deviations 

Machine generated alternative text:
13.57. 
13.9. 
950/0 

*Formula:*

Machine generated alternative text:
Formula to find population mean 
Ex 
n 
Formula to find population standard deviation 
n 
Formula to find the z-score 
z score 

Formula to find the z-score when we have multiple samples:

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image004.png

*Example Problem:*

Machine generated alternative text:
The test scores for a class are normally distributed. The test scores for a class are normally distributed. 
Given: and 0=10 
Given: and 0=10 
What is the probability a student scored above@What is the probability a student scored above 60 
83.31 % 
0-7S 
10 

*Problem to find out z-score when sample is given:*

Machine generated alternative text:
Sample problem: In general, the mean height of women is 65" with a standard deviation of 3.5". What is the 
probability of finding a random sample of 50 women with a mean height of 70", assuming the heights are normally 
distributed? 

Machine generated alternative text:
= (70 - 65) / (3.5"/50) = 5/0.495 = 10.1 
The key here is that we're dealing with a sampling distribution of means, so we know we have to include the 
standard error in the formula. We also know that 99% of values fall within 3 standard deviations from the mean in 
a normal probability distribution (see 68 95 99.7 rule). Therefore, there's Less than 1% probability that any sample 
of women will have a mean height of 70" 

**[2] t-score**

Machine generated alternative text:
The t score formula is: 

When to use t-score?

* Has a sample size below 30,
* Has an unknown population standard deviation.

Machine generated alternative text:
DO you know the population 
standard deviation, o, ? 
Yes 
Is the sample size 
aoove 30? 
Yes 
No 
use the z-score 
No 
use the t-score 
use the t-score 
• Replace s in the t-score formula with c 

Machine generated alternative text:
The average test score of a population is 75. A sample 
of 9 students are randomly selected. The standard 
deviation for the sample is 10. What is the probability 
the average score for the sample is above 80? 
n: q s: 10 
T Value 
= 1.5 
Degrees of Freedom (df) = 8 
Calculate 
Two tailed P- Value = 0.172003 
One tailed P- Value = 0.036002 
pc k 780): 
: .oto 

*When to use z-score and when to use t-score? Examples*

Machine generated alternative text:
The average test score for an entire school is 75 The average test score for an entire school is 75. 
The standard deviation of a random sample of 
with a standard deviation of 10. What IS the 
probability that a random sample of 5 students students is p. What is the probabilty the 
scored above 80? 
Conditions for usinq t 
1. 
2. n<30 
The average test score for an entire school is 75. 
The sgndatd-dgua.t.lm of a ranüG-GöiÖÖf9 
students is ID. What is the probabilty the 
average test score for the sample is above 80? 
average test score for the sample is above 80? 
Conditions for usinq t 
l.c is unknown 
Conditions for usinq t 
l.c is unknown l/ 
n: 

**[3] Central Limit Theorem**

* Mean of the samples frequencies gives the normal distribution irrespective of the nature of population data.

i.e. when the population is not normally distributed, we shall pick samples from the population, compute the mean of it and plot it in a separate graph. When the number of sampling increases, sample distribution will approach the normal distribution.

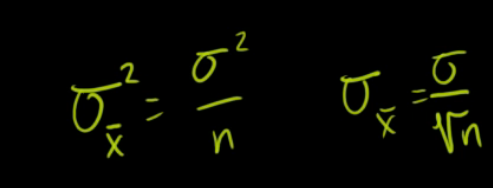
* Suggested sample size >= 30. Suggested number of samples = infinity (∞)
* Approximately correct probability can be computed from the sampling distributions.
* When the size of N increases standard deviation will come close to mean of sample distribution.

Machine generated alternative text:
mean= 
median= 
sd= 
skev.= 
Reps= 
mean= 
median= 
sd= 
skev.= 
Reps= 
mean= 
median= 
sd= 
skev.= 
13.92 
12.00 
s_6S 
0.42 
_0_99 
100000 
13.92 
14.00 
2.73 
0.13 
-0.04 
100000 
13.91 
14.00 
1.73 
cos 
0.13 
Parent population (can be changed with the mouse) 
Sample Data 
Distribution of Means: N 
Distribution of Means: N=2S 
Clear lower 3 
Normal , 
Sample: 
Animated 
5 
10,000 
100,000 
Mean 
o 
Fit normal 
Mean 
N=25 
o 
Fit normal 

Sampling distribution mean is same as population mean

Machine generated alternative text:
Sam p e 
yueø 

Standard Error of the sampling mean:



**Sample Problem:**

Machine generated alternative text:
The average male drinks of water when active outdoors (with 
a standard deviation of(L!- . You are planning a full day nature 
trip for 50 men and will bring 110 L of water. What is the 
probability that you will run out? 
00k) 
v.aore 
o 
50 
2 
(3.4 
.02-1? 
o.oqq 
i la z .02 0 4) . 
So it's a 2.17% chance 
we run out of water. 

**[4] Hypothesis Testing**

*Hypothesis* : is a claim that we want to test.

Null Hypothesis - Ho - Currently accepted value for a parameter.

*Alternative Hypothesis*- Ha - Also called research hypothesis. Involves the claim to be tested.

*Level of Confidence:* How confident are we in our decision ? It is referred with letter "C".

e.g. C-95%, C-99%

*Level of Significance:* α = 1-C

***Steps Involved:***

***(i) Set up Null (Ho) and Alternative Hypothesis(Ha)***

we need to test the difference between the sample and population mean due to a random chance.

Ho - there is no difference between sample and population.

Ha - Assumption on Null hypothesis is wrong.

***(ii) Set the Criteria for decision:***

- it refers the level of significance for a test and it will differ based on the business requirements.

- Based on the significance level, we may accept or reject Null hypothesis.

(iii) Compute the random chance of probability

- Random chance of probability helps to determine the higher probability and evidences to accept Null hypothesis.

(iv) Make decision

- compare the p value with the predefined significance level.

- if its less than the significance level, reject Null hypothesis else accept it.

- there are four possible decisions:

(a) The decision to retain the null hypothesis could be correct.

(b) The decision to retain the null hypothesis could be incorrect, it is know as **Type II error**.

( c) The decision to reject the null hypothesis could be correct.

(d) The decision to reject the null hypothesis could be incorrect, it is known as **Type I error**.

**Example Problem:**

###### *Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive effect on blood glucose levels. A sample of 36 patients who have tried the raw cornstarch diet have a mean glucose level of 108. Test the hypothesis that the raw cornstarch had an effect or not.*

Step 1: State the hypothesis.

Population mean is 100

Ho : μ = 100

Ha : μ > 100

Step 2: Set up the significance level

As there is no significance level given, consider the default significance level 5% (0.05).

Step 3: Compute the random chance of probability using z-score (and z table)

Machine generated alternative text:
(T 
Mean 
Standard Deviation 
For this set of data: (108-100) / 

**p-value associated with 3.20 is 0.9993.** i.e. probability of having value less than 108 is 0.9993 and more than or equals to 108 is (1-0.9993)=0.0007.

Step-4: It is less than 0.05 so we will reject the Null hypothesis i.e. there is raw cornstarch effect.

***Non-Directional (or Two Tail) hypothesis testing:***

In one tail test, we reject the Null hypothesis if the sample mean is either positive or negative extreme any one of them. But, in case of two tail test we can reject the Null hypothesis in any direction (positive or negative).

Two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha in the other direction. This means that .025 is in each tail of the distribution of your test statistic. Why are we saying 0.025 on both side because normal distribution is symmetric. Now we come to a conclusion that Rejection criteria for Null hypothesis in two tailed test is 0.025 and it is lower than 0.05 i.e. two tail test has more strict criteria to reject the Null Hypothesis.

Machine generated alternative text:
p s 005 
one-tail cr 
•cal region 
p s 0.025 
two-tail critical region(s) 

**Example Problem**

Machine generated alternative text:
Templer and Tomeo (2002) reported that the population mean score on the quantitative portion of the Graduate 
Record Examination (GRE) General Test for students taking the exam between 1994 and 1997 was 558 ± 139 (p 
± o). Suppose we select a sample of 100 participants (n = 100). We record a sample mean equal to 585 (M = 
585). Compute the p-value to check whether or not we will retain the null hypothesis (p = 558) at 0.05 level of 
significance (a = .05). 
Solution: 
Step-I : State the hypotheses. The population mean is 558. 
HO: 558 
HI: p 558 (two tail test) 
Step-2: Set up the significance level. As stated in the question, it as 5% (0.05). In a non-directional two-tailed 
test, we divide the alpha value in half so that an equal proportion of area is placed in the upper and lower tail. So, 
the significance level on either side is calculated as: a/2 = 0.025. and z score associated with this (1- 
0.025=0.975) is 1.96. As this is a two-tailed test, z-score(observed) which is less than -1.96 or greater than 1.96 
is a evidence to reject the Null hypothesis. 

Machine generated alternative text:
For this set of data: (585-558) / 
You can look at the probability by looking at z- table and p-value associated with 1.94 is 0.9738 i.e. probability o 
having value less than 585 is 0.9738 and more than or equals to 585 is (1-0.9738)=0.03 
Step-4: Here, to make a decision, we compare the obtained z value to the critical values (+1- 1.96). We reject the 
null hypothesis if the obtained value exceeds a critical values. Here obtained value (ZObt= 1.94) is less than thc 
critical value. It does not fall in the rejection region. The decision is to retain the null hypothesis. 
The obtained value is 1.94, 
which fails to reæh the cutoff 
for rejection regm, retain 
the null hypottwsis. 
Rejection region 
.0250 
Rejection region 
.0250 
-a 
-2 
-1.96 
Retain the null 
hypdhesis 
Null 
2 
3 

**[5] ANOVA**

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of three or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

**Formula Sequence:**

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image020.png

(2) Calculate Sum of Squares of Total value (SS Total)

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image021.png

(3) Calculate SS *between* value

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image022.png

(4) Calculate SS within value

**SSwithin = SS Total - SS between**

(5) Calculate MS between value

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image023.png

(6) Calculate MS within value

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image024.png

(7) Calculate F value

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image025.png

Machine generated alternative text:
correlation factor 
number of elements from all the groups 
number of elements within the group 
k 
ANOVA 
Between 
Within 
Total 
number of groups 
df (deg. Of freedom) 
k-l 
SS (sum of squares) 
(Ex)2 
SSwithin = SS Total - SS between 
SS Total 
MS (mean squares) 
SS between 
MS between = 
df between 
SS within 
MS within= 
df within 
MS 
b e tween 
MSW i t h in 

Problem solving - steps

1. Define Null and Alternative Hypotheses

2. State Alpha

3. Calculate Degrees of Freedom

4. State Decision Rule

5. Calculate Test Statistic

6. State Results

7. State Conclusion

Sample Problem

Researchers want to test the reaction of new medication. They split participants into three conditions (0mg, 50mg, and 100mg), then ask them to rate their anxiety level on a scale of 1-10. Are there any differences between the three conditions using alpha = 0.05?

Machine generated alternative text:
Ex 
mg 
57 
50mg 
47 
IOOmg 
21 

1. Null and Alternative Hypothesis

Ho : μ0mg = μ50mg = μ100mg

H1 : not all μ are equal

1. Alpha selection

Alpha (α ) = 0.5

1. Degrees of freedom

N = 21 , n = 7, k = 3 (no. of groups )

df between = 3-1=2 (k-1)

df within = 21-3=18 (N-k)

df total = 21-1 =20 (N-1)

Now look at the F-table and find the critical value using (2,18) [ i.e. df between and df within] and alpha = 0.05. The result is **3.5546 (critical value )**

(4) Decision Rule

So the decision rule is : if the calculated F value is greater than **3.5546,** null hypothesis should be rejected.

(5) Calculate Test Statistic

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image028.png

Machine generated alternative text:
mg 
OmgA2 
Omg 
81 
49 
81 
467 
47 
50mgA2 
49 
36 
36 
49 
49 
36 
319 
00mg 
21 
IOOmgA2 
16 
16 
67 

C:\0F3190E5\D9A26DA1-B572-42DA-ACC5-5CAF44FF1ABF_files\image030.png

=> (464.14 + 315.58 + 63) - 744.05 = **98.67**

SS total = (467 + 319 + 67 ) - 744.05 = **108.95**

SS within = 108.95 - 98.67 = 10.28

MS between = 98.67 / 2 = **49.335**

MS within = 10.28 / 18 = **0.5711**

(6) Calculate Result (F )

F = 49.34/0.57=**86.5614**

Result = Reject Null hypothesis

(7) Conclusion

three conditions differed significantly on anxiety level F(2, 18) = 86.56 , p < 0.05

[6] Correlation coefficient

- shows the linear relationship between two sets of data.

Machine generated alternative text:
n(Exy) - 

References

[1] <https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/>

[2] AOVA

<http://www.statisticslectures.com/topics/onewayanova/>

<https://www.kean.edu/~fosborne/bstat/08ANOVA.html>