Hypothesis Testing II: Z and T test

Data Science Immersive

May 29, 2019

FLATIRON SCHOOL

Agenda today

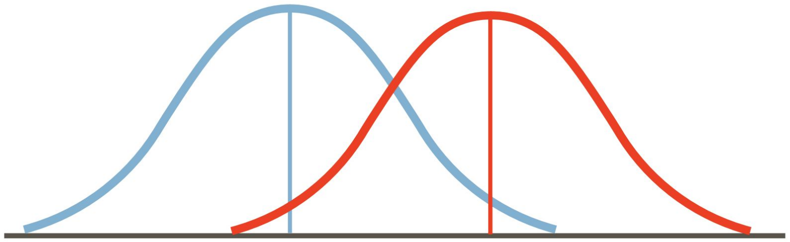
- Review z score and p values
- Review of theory of hypothesis testing
- Practice hypothesis testing with z and t tests
- Introduction to error and confidence interval
- Interpreting results of hypothesis testing, confidence interval, and errors

SWBAT

- Formulate scientific hypothesis and identify the right test statistics
- Perform z test and t tests
- Make the correct statistical decision and interpret the results

Let's review: z score

- A z distribution is known as the standard normal distribution with a mean of 0 and a sd of 1
- A z score represents how many sd's you are away from the mean
- We will be able to compute a raw score from a z score, and vice versa



Let’s practice: hypothesis testing with z test

- If we have a sample size of >100, or known population standard deviation σ, we can perform the hypothesis testing with z test
- Steps for hypothesis testing with z:
  - Formulate the problem
  - Operationalize the problem and identify the data
  - Identify test statistics
  - Choose the alpha
  - Calculate the test statistics
  - Compare the calculated z score to the **z critical values**
  - Make a decision
  - Interpret the results

$$\text{Z Score} = \frac{x - \mu}{\sigma} = \frac{\text{Raw score - Mean}}{\text{Standard deviation}}$$

Let’s practice: hypothesis testing with z test

- Problem formulation
  - Is our class significantly different from the general population in egomaniac score?

Let’s practice: hypothesis testing with z test

- Operationalize the problem and identify the data
  - Egomaniac score is normally distributed in the population with a $\mu = 50$, sd = 5
  - Our class has the following statistics
    - Mean of 37.64
    - Sd of 12.87

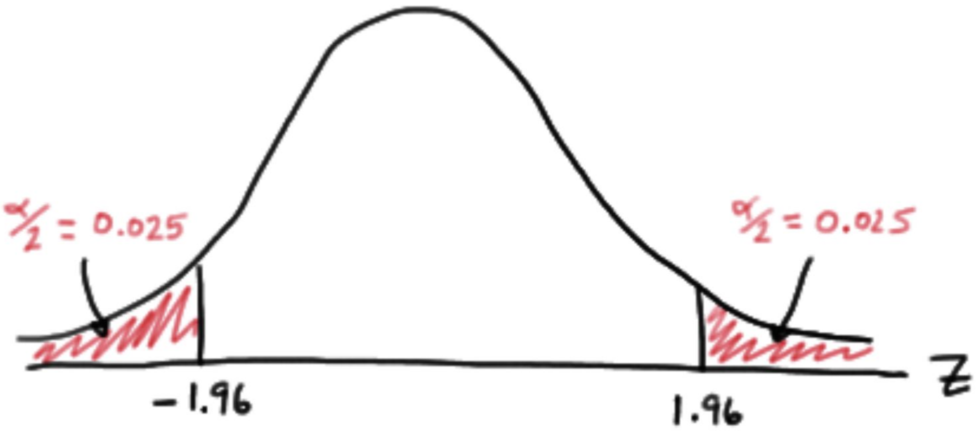
Scores
45%
26%
50%
31%
26%
29%
76%
42%
31%
27%
39
29%
25%
49%
36
35%
44%

Let’s practice: hypothesis testing with z test

- Identify the test statistics
  - Since the population SD is known, we can use the z test
  - $$\text{Z Score} = \frac{x - \mu}{\sigma} = \frac{\text{Raw score - Mean}}{\text{Standard deviation}}$$

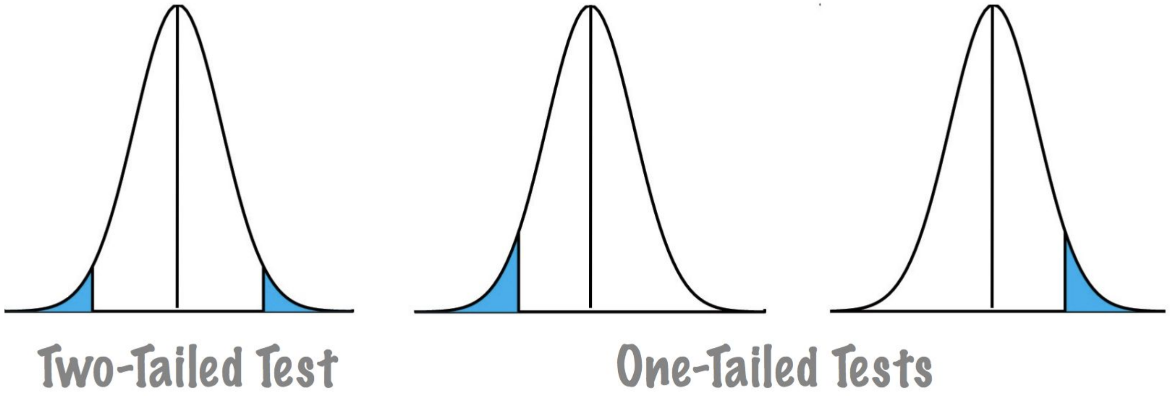
Let's practice: hypothesis testing with z test

- Choose the alpha (region of rejection, or risks we are willing to take to make a false positive)
  - Alpha = 0.05



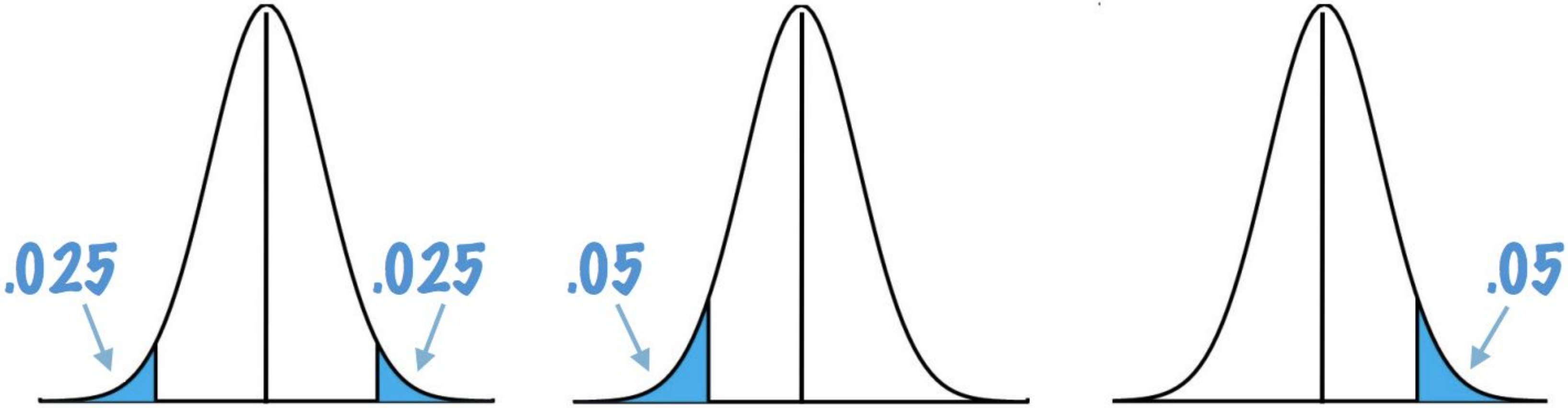
Let's practice: hypothesis testing with z test

- A note on one vs two-tailed tests



One-tailed tests are easier to reach statistical significance in the predicted tail, but rule out the possibility of testing the other tail

Let’s practice: hypothesis testing with z test

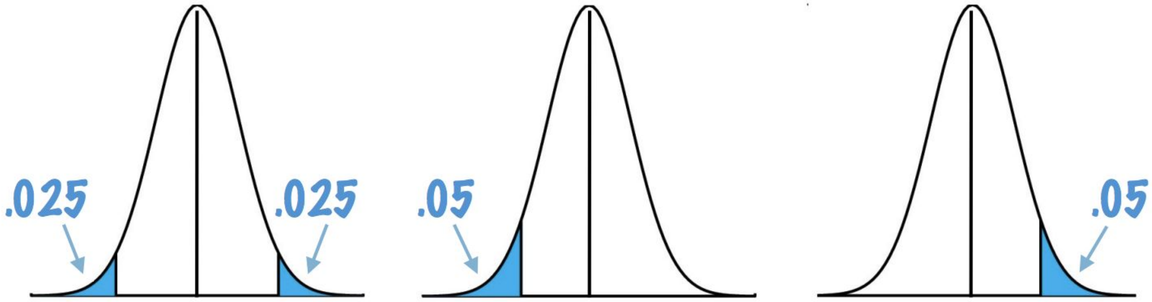


Let’s practice: hypothesis testing with z test

- Calculate the test statistics:
  - $$Z\ Score = \frac{x - \mu}{\sigma} = \frac{\text{Raw score - Mean}}{\text{Standard deviation}}$$

Let’s practice: hypothesis testing with z test

- Compare the calculated z score to the **z critical values**



- The rejection region is the area of the normal distribution in which we reject the null hypothesis
- The z-scores that serve as the boundary of this area are called critical z-scores (zcrit) One-tailed test: zcrit = +/- 1.645 Two-tailed test: zcrit = -1.96 and +1.96

Let’s practice: hypothesis testing with z test

- Compare the calculated z score to the **z critical values**
  - If our calculated z values is **greater** than the z critical values, we say that we reject the null
  - The corresponding p values acquired is less than 0.05
  - This means that our results are so extreme that this difference cannot be due to chance

Let’s practice: hypothesis testing with z test

- Make a decision
  - The results are statistically significant and you can reject H0 when: **p < .05** and **zcalc > zcrit**
  - How does the z scores compare?
  - How does the p value compare?
  - What is your decision?

Let’s practice: hypothesis testing with z test

- Interpret the results
  - Since our calculated z score z = is than critical z value with a p = .0068, we therefore the null hypothesis and conclude that our results . In other words, our class than the general population.

Errors

- Just because we have finished calculated the test statistics, our jobs aren’t done. We need to be mindful of the possible errors we might make
    - Type I error: false positive
    - Type II error: false negative

Errors

- Type I Error (α)
  - Rejecting the null hypothesis when it is in fact true
  - i.e., saying there is a difference when there really isn’t
- Type II Error (β)
  - Failing to reject the null hypothesis when it is in fact false
  - i.e., saying that there is no difference when there really is

Let’s practice: hypothesis testing with t tests

- Sometimes when the population sd is unknown and we have a small sample size, we need to use the t test
- Different types of t tests:
  - **One sample t test**: mean of single group against a population mean
  - **Independent sample t test**: how do the means of two independent groups compare?
  - **paired sample t test**: how does the mean of the same group compare between different times?

Let’s practice: hypothesis testing with t tests

- Example: is our class significantly different from the population in terms of humor?
  - We know that humor is normally distributed with a mean of 50
  - Our class has a mean of 58.89 and a sd of 4.56, and a sample size of $n = 18$

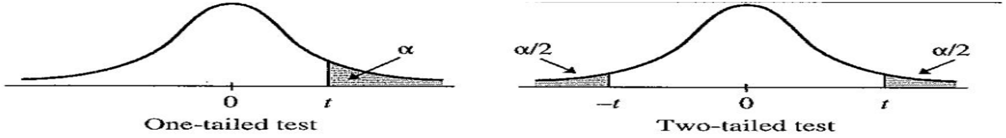
Let’s practice: hypothesis testing with t tests

- Since we no longer have knowledge of the population standard deviation, we need to **estimate** it using the standard error
- What is the standard error?

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

Let’s practice: hypothesis testing with t tests



	Level of Significance for One-Tailed Test					
	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test					
df	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.620
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Let’s practice: hypothesis testing with t tests

- Calculate the test statistics for our question
  - What null hypothesis is this?
  - What type of t test is it?
  - What is the value of our t score?
  - What decision have we made?
  - How do you interpret the results?

Let's practice: hypothesis testing with t tests

- Example 2: you are curious to see whether sertraline or escitalopram is more correlated with increased mood for depressed patients. Assuming there is no significant difference in self-reported mood before treatment, you are comparing their self-reported mood after treatment. You found:
  - Mean of zoloft users is 7.4 with a sd of 0.5, n = 14
  - Mean of Lexapro users is 6.9 with a sd of 1.2, n = 17
  - df = n1 + n2 - 2

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

Let’s practice: hypothesis testing with t test

What is our null hypothesis?

- $\mu 1 = \mu 2$, or $\mu1-\mu2 = 0$
- There is no mean difference between zoloft users and lexapro users in perceived mood after treatment.

What is our alternative hypothesis?

- $\mu 1$ =/= $\mu 2$, or $\mu1-\mu2$ =/= 0
- There is significant mean difference between zoloft users and lexapro users in perceived mood after treatment.

Let's practice: hypothesis testing with t test

- Homogeneity of variances
  - If one group's variance is no more than 2x the other group, we can assume homogeneity of variance
  - It essentially is assuming that two groups don't differ in variability

Let’s practice: hypothesis testing with t test

- To decide which t test to use, use the flow chart

1) Are both sample sizes large?	**Yes**: large sample test for independent means **Formula 7.3**	$z = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
	No: (small sample sizes)	Go to question #2
2) Are the sample sizes equal?	**Yes**: pooled variances test for equal sample sizes **Formula 7.8**	$t = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{s_1^2 + s_2^2}{n}\right)}}$
	No: (unequal sample sizes)	Check for homogeneity of variance[+]: Go to question #3
3) Can the population variances be assumed equal?	**Yes**: pooled variances test **Formula 7.5B**	$t = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$
	No (variances unequal):	$t = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{}$

Let’s practice: hypothesis testing with t test

- Calculate the standard error of difference
    - this is the standard deviation for the distribution of difference scores

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Let’s practice: hypothesis testing with t test

- Can we assume homogeneity of variance?
- Which test should we use?

Let’s practice: hypothesis testing with t test

- Formulate the hypothesis
- Calculate the test statistics
- Compare the calculated t to the t critical values
- Make conclusions
- What error could we possibly have made?

Let's practice: hypothesis testing with t test

- Calculating t tests using scipy

ttest_ind underestimates p for unequal variances:

```
>>> rvs3 = stats.norm.rvs(loc=5, scale=20, size=500)
>>> stats.ttest_ind(rvs1, rvs3)
(-0.46580283298287162, 0.64145827413436174)
>>> stats.ttest_ind(rvs1, rvs3, equal_var = False)
(-0.46580283298287162, 0.64149646246569292)
```

When n1 != n2, the equal variance t-statistic is no longer equal to the unequal variance t-statistic:

```
>>> rvs4 = stats.norm.rvs(loc=5, scale=20, size=100)
>>> stats.ttest_ind(rvs1, rvs4)
(-0.99882539442782481, 0.3182832709103896)
>>> stats.ttest_ind(rvs1, rvs4, equal_var = False)
(-0.69712570584654099, 0.48716927725402048)
```

T-test with different means, variance, and n:

```
>>> rvs5 = stats.norm.rvs(loc=8, scale=20, size=100)
>>> stats.ttest_ind(rvs1, rvs5)
(-1.4679669854490653, 0.14263895620529152)
>>> stats.ttest_ind(rvs1, rvs5, equal_var = False)
(-0.94365973617132992, 0.34744170334794122)
```

Confidence Interval

- Instead of making a point estimate, we can estimate **an interval** in which the mean value may fall in
- We frame the width of the interval as the level of confidence in which we believe the mean may fall into
- It is calculated as

$$\mu = \overline{X} \pm t_{crit} s_{\overline{X}}$$

Confidence Interval

- The value of confidence interval is affected by:
    - The standard error
    - The sample mean
    - Alpha -> how does alpha affect the width of confidence interval?

- The confidence interval can be used for
    - Estimation of mean
    - **Hypothesis testing**

Let's practice: hypothesis testing with t test

- Calculate Confidence Interval with our previous examples
  - One sample t test with humor score
  - Two sample independent test with different antidepressants

$$\mu = \overline{X} \pm t_{crit} s_{\overline{X}}$$

$$\mu = \bar{X} \pm t_{crit} s_{\bar{x}}$$

$$\boxed{\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{crit} s_{\bar{X}_1 - \bar{X}_2}}$$

Confidence Intervals

- Confidence interval for alpha = .05
  - Humor score confidence interval
  - Antidepressants confidence interval
- Confidence interval for alpha = .01
  - Humor score confidence interval
  - Antidepressant confidence interval

$$\mu = \overline{X} \pm t_{crit} s_{\overline{X}}$$

Confidence Intervals

- Interpretation of the confidence interval:
  - We are 95% confident that the population mean for humor score falls between ___ and ___.
  - We are __% confident that the population mean difference of self-reported mood score falls from ___ and ___.

Conclusions

- Today you learned how to perform z tests, one sample t test, and two sample independent t tests, and calculate confidence interval
- Tomorrow we will cover
  - Power analysis and effect size
  - Comparing multiple groups - analysis of variance (ANOVA)