SGA - Basic Statistics - Isupov Ilya

May 2024

Telegram: @Jeremix

1 Hypothesis testing meets confidence intervals

There is a connection between confidence intervals and hypothesis testing. Assume that we have an i.i.d. sample $x = (x_1, \dots, x_n)$ from some random variable X with finite variance. Consider one-sample t-test with null hypothesis $EX = \mu_0$ and symmetric alternative. For simplicity, let us assume that n is large enough and replace T-distribution with standard normal distribution. Assume that one found confidence interval I for EX with confidence level 95%. Prove that standard decision-making procedure of t-test is equivalent to the following: reject null hypothesis if and only if μ_0 does not belong to I. Follow the plan.

$$H_0 : EX = \mu_0$$

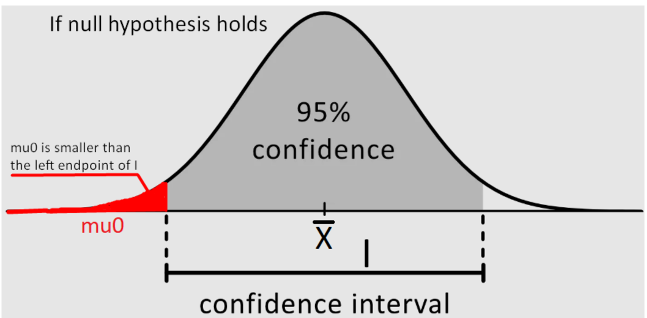
$$H_1 : EX \neq \mu_0$$

1. Assume that null hypothesis holds. We believe that t-statistics in this case is distributed according to standard normal law (due to assumption that n is large). Recall that t-statistics for sample x is defined as:

$$t \approx \frac{\bar{x} - \mu_0}{SD(x)} \cdot \sqrt{n}$$

2. If μ_0 does not lie in I, either μ_0 is larger than the right endpoint of I or μ_0 is smaller than the left endpoint of I. Let us consider the latter case.

3. Consider event "μ_0 is smaller than the left endpoint of I". Write this condition as an inequality using μ_0, $\bar{x}$, $SD(x)$, n and a constant 1.96. (Recall that we assume that null hypothesis holds.)



$$s = 1.96 \cdot \frac{SD(x)}{\sqrt{n}}$$

$$\mu_0 < \bar{x} - s$$

$$\mu_0 < \bar{x} - 1.96 \cdot \frac{SD(x)}{\sqrt{n}}$$

4. Transform this inequality such that it becomes $(\dots) > 1.96$. Does the left-hand part look similar to something?

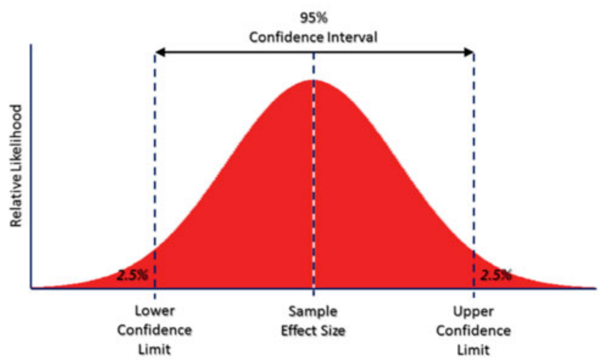
$$\frac{\bar{x} - \mu_0}{\frac{SD(x)}{\sqrt{n}}} > 1.96$$

The left-hand part of this formula looks similar to t-score (t-statistic):

$$t(x) = \frac{\bar{x} - \mu_0}{\frac{SD(x)}{\sqrt{n}}}$$

5. Recall why we use number 1.96, how it is connected to standard normal distribution.

If we consider the standard normal distribution $N(0,1)$, we will find that 95% of the values lie in the range $[-1.96, 1.96]$. The total area of the tails not included in the 95% interval $[-1.96, 1.96]$ is the remaining 5%, 2.5% for each tail. In other words we can say that 1.96 is 97.5 percentile in the standard normal distribution.



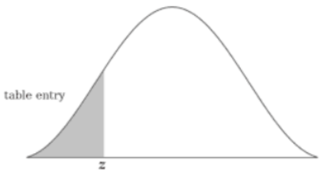
6. Find probability that μ_0 is smaller than the left endpoint of I provided that null hypothesis holds.

$$P(\mu_0 < \text{left endpoint of I } |H_0) = P(\mu_0 < \bar{x} - 1.96 \cdot \frac{SD(x)}{\sqrt{n}})$$

Let's use Z-table and find this probability:

$$P(\mu_0 < \text{left endpoint of I } |H_0) = 0.025 = 2.5\%$$

Negative Z score table



Use the negative Z score table below to find values on the left of the mean as can be seen in the graph alongside. Corresponding values which are less than the mean are marked with a negative score in the z-table and respresent the area under the bell curve to the left of z.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330

7. Find probability that μ_0 does not lie in I provided that null hypothesis holds.
μ_0 does not lie in I if:
- μ_0 is smaller than the left endpoint of I:

$$P(\mu_0 < \text{left endpoint of I } |H_0) = 0.025 = 2.5\%$$

- μ_0 is larger than the right endpoint of I
Let's find the probability that μ_0 is larger than the right endpoint of I using Z-table:

$$P(\mu_0 > \text{right endpoint of I } |H_0) = P(\mu_0 > \bar{x} + 1.96 \cdot \frac{SD(x)}{\sqrt{n}}) = 0.025 = 2.5\%$$

Probability that μ_0 does not lie in I provided that null hypothesis holds:

$$P(\mu_0 \text{ not in I } |H_0) = P(\mu_0 < \text{left endpoint of I } |H_0) + P(\mu_0 > \text{right endpoint of I } |H_0) =$$

$$= 2.5\% + 2.5\% = 5\% = 0.05$$

8. Assume we are following rule "reject null hypothesis if and only if μ_0 does not belong to I." Find probability that we falsely reject null hypothesis provided that it is true.
It's Type I error - reject H_0 when H_0 is true.

$$P(\text{reject } H_0|H_0 \text{ is true}) = 0.05 = 5\%$$

9. Explain in what cases (in terms of $\bar{x}$) will we reject null hypothesis if we follow mentioned rule.
We will reject H_0 if $\bar{x}$ is very different from μ_0:

$$\bar{x} > \mu_0 + 1.96 \cdot \frac{SD(x)}{\sqrt{n}}$$

$$\bar{x} < \mu_0 - 1.96 \cdot \frac{SD(x)}{\sqrt{n}}$$

It means that if the value of t-statistics don't fall into the interval $[-1.96; 1.96]$ - we will reject null hypothesis.

10. Explain that this rule is equivalent to the rule used in ordinary two-sided one-sample t-test.

This rule is equivalent to the rule used in ordinary two-sided one-sample t-test because:

1. In the case of two-sided one-sample t-test, we reject the null hypothesis if the value of the t-statistic goes beyond the critical region - in other words if $|\frac{\bar{x}-\mu_0}{\frac{SD(x)}{\sqrt{n}}}| > 1.96$ (provided that n is large enough).

2. In the case of 95% confidence interval we also reject null hypothesis if $|\frac{\bar{x}-\mu_0}{\frac{SD(x)}{\sqrt{n}}}| > 1.96$.