RBCDSAI IITM

Mathematical Foundations of Data Science

Session 5 – Inferential Statistics – 2

Nandan Sudarsanam,
Department of Data Science and AI,
Wadhwani School of Data Science and AI,
Indian Institute of Technology Madras

Introduction to hypothesis testing

- Inferential statistics
  - Given a sample statistic ($\bar{x}$) what can I say about the population parameter (μ): Confidence Intervals
  - Given a sample can I answer pointed questions about the population parameters: Hypothesis Tests
- How are these related? What does one tell you about the other?

The general rubric

- Have a null and alternate hypothesis (Mutually Exclusive Collectively Exhaustive)
- Do some basic calculations/arithmetic on the data to create a single number called the “test statistic”
- If we assume the null hypothesis to be true (and make some assumptions about the distributions of various variables), then the ‘test statistic’ should be no different than a single random draw from a specific probability distribution.
- Ascertain the probability of observing a “test statistic” equal to or more extreme than that which was observed, from this theoretical distribution, assuming that the null hypothesis is true. This is the *p*-value.
- Reject the null hypothesis if the *p*-value is low
- Ergo: It’s P(Data|Hypothesis) not P(Hypothesis|Data)

Further Discussion

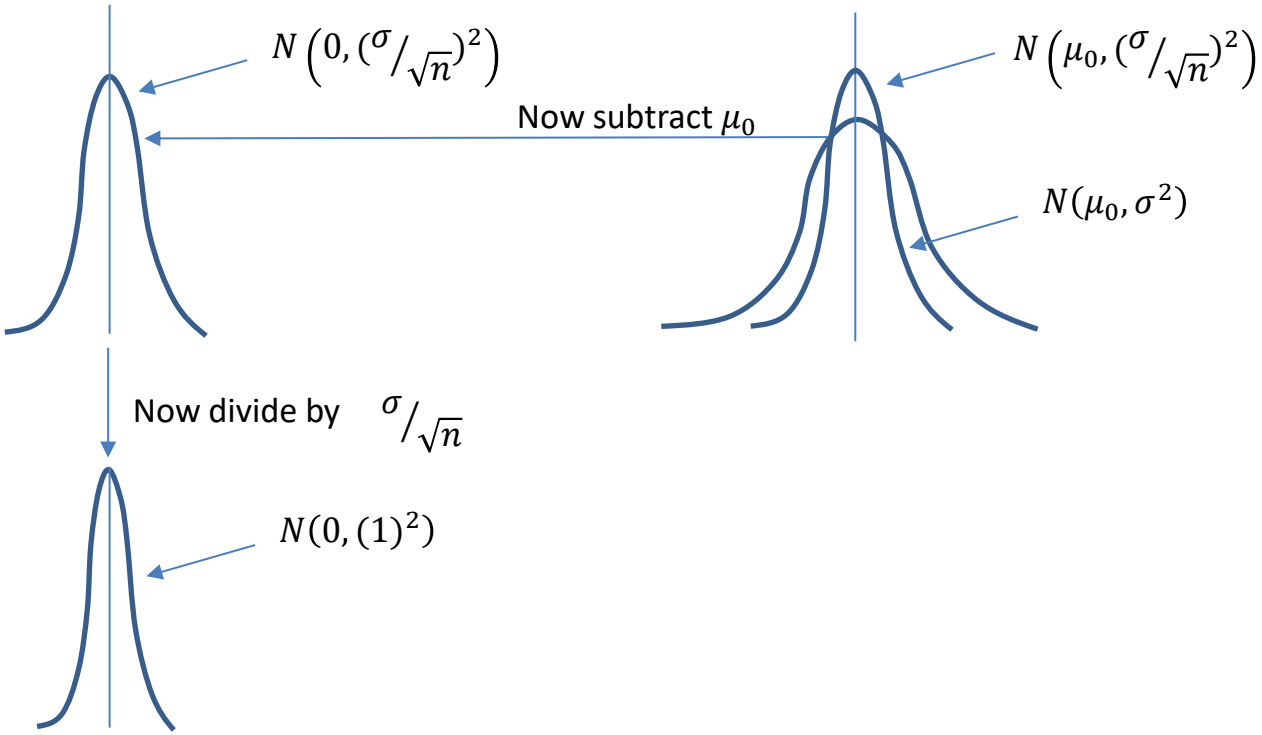
- What is α in all of this?
- What is β? We will come back to this.
- What does probability of observing *a test statistic more extreme than that which was observed* mean? How do we define what is likely and what is unlikely?
- Why have I heard other definitions? Like: “The *P*-value is the smallest level of significance that would lead to rejection of the null hypothesis $H_0$with the given data.”

The general rubric

- Have a null and alternate hypothesis (Mutually Exclusive Collectively Exhaustive) $H_0: \mu \leq 210$ and $H_{alt}: \mu > 210$
- Do some basic calculations/arithmetic on the data to create a single number called the "test statistic" $z_{stat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
- If we assume the null hypothesis to be true (and make some assumptions about the distributions of various variables), then the 'test statistic' should be no different than a single random draw from a specific probability distribution. This is the Z-distribution or $N(0,1^2)$
- Ascertain the probability of observing a "test statistic" equal to or more extreme than that which was observed, from this theoretical distribution, assuming that the null hypothesis is true. This is the *p*-value. Use Z-tables, any other software
- Reject the null hypothesis if the *p*-value is low

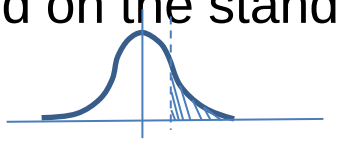
Single sample z-test

- Why is the z-stat a z distribution (or in other words a $N(0, 1^2)$)

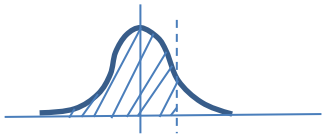


Single sample z-test

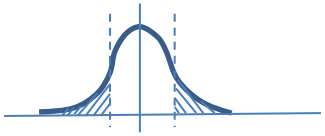
- The p-value is the probability of seeing a test statistic as extreme as the calculated value if the null hypothesis is true. This based on the notion of what values are unlikely.
- If z_{stat} was computed to be 1.2 then
- P-value, Based on the standard null hypothesis:
- H_0: $\mu \leq 210$



- If null hypothesis was:
- H_0: $\mu \geq 210$



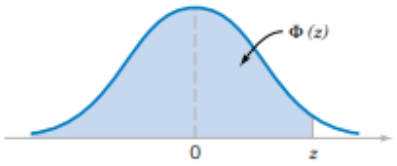
- H_0: $\mu = 210$ (two tailed)



Sample tables and Excel functions

- Using Z-tables:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$



z	0.00	0.01	0.02
0.0	0.500000	0.503989	0.507978
0.1	0.539828	0.543795	0.547758
0.2	0.579260	0.583166	0.587064
0.3	0.617911	0.621719	0.625516
0.4	0.655422	0.659097	0.662757
0.5	0.691462	0.694974	0.698468
0.6	0.725747	0.729069	0.732371
0.7	0.758036	0.761148	0.764238
0.8	0.788145	0.791030	0.793892
0.9	0.815940	0.818589	0.821214
1.0	0.841345	0.843752	0.846136
1.1	0.864334	0.866500	0.868643
1.2	0.884930	0.886860	0.888767

- Using Excel: =norm.s.dist(1.2, TRUE)
- Using R: > pnorm(1.2)

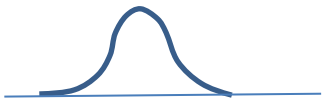
Solving a problem end-to-end

- We historically receive an average of 5000 visitors to our web page each day. A new digital Ad claims to increase the flow of visitors by at least 2000. We know that the standard deviation of visitors is 200. We run the digital Ad for 10 days and collect this data:

  No. of visitors in each day: 6719, 6834, 6808, 7092, 7100, 7011, 6744, 6933, 6809, 6760

  What conclusion can we make of the claim?

- what is the null hypothesis: $H_0: \mu \geq 7000$
- What is the test statistic: $\frac{6881 - 7000}{200/\sqrt{10}} = -1.881$
- Which side of the z-distribution are we interested in?
- What is the p-value? 0.0299



What else can we hypothesize

One –sample tests	Test statistic	Distribution used	Examples
z-test	$z=\frac{\bar{x}-\mu_0}{(\sigma/\sqrt{n})}$	N(0,1^2) or Z	Already discussed
t-test	$t=\frac{\bar{x}-\mu_0}{(s/\sqrt{n})}$	t_{n-1}	Same as z but standard deviation unknown
Chi-Square test	$\chi^2=(N-1)\frac{s^2}{\sigma_0^2}$	χ^2_{n-1}	Test for variance
z-proportion test	$z=\frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n}$	N(0,1^2) or Z	Testing a proportion

What if we want to make inferences concerning two populations

- Examples:
  - Dropping two different balls to see if one takes more time on average?
  - Two different agents are making sales calls, which one is better?
  - We have two machines and we have confirmed that both are on average producing bags of 100gms. Do they both have the same variability?
- Single set of data versus two sets of data
- Think of it as dealing with two variables for the first time

The general rubric

- Have a null and alternate hypothesis (Mutually Exclusive Collectively Exhaustive) $H_0: \mu_1 = \mu_2$ and $H_{alt}: \mu_1 \neq \mu_2$
- Do some basic calculations/arithmetic on the data to create a single number called the "test statistic" $z = \frac{(\overline{x_1} - \overline{x_2}) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- If we assume the null hypothesis to be true (and make some assumptions about the distributions of various variables), then the 'test statistic' should be no different than a single random draw from a specific probability distribution. This is the Z-distribution or $N(0,1^2)$
- Ascertain the probability of observing a "test statistic" equal to or more extreme than that which was observed, from this theoretical distribution, assuming that the null hypothesis is true. This is the *p*-value. Use Z-tables, any other software
- Reject the null hypothesis if the *p*-value is low

Some common cases for hypothesis tests

Two –sample tests	Test statistic	Distribution	Examples
z-test	$z = \frac{(\overline{x_1} - \overline{x_2}) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z or $N(0,1^2)$	Any comparison of mean with known variance
t-test (independent, equal variance)	$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$t_{n1+n2-2}$	Two different settings of the same machine, with tests on separate experimental units?
t-test (independent, unequal variance)	$t = \frac{(\overline{x_1} - \overline{x_2}) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t_k ,where $k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$	Do Mac users spend more than PC Users on Average?
Paired t-test	$t = \frac{\overline{di} - d_0}{(\frac{s_{di}}{\sqrt{n}})}$	t_{n-1}	Pre-post on the same experimental unit; related rows
f-test	$F = \frac{s_1^2}{s_2^2}$	$F_{n1,n2}$	Compare two population variances
z-proportion test	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$	Z or $N(0,1^2)$	Are Mac users more likely to buy than PC Users?

CI and tests equivalency

- We saw 4 CIs that involved a single sample and we saw 4 equivalent hypothesis tests
- What does a CI means for the two-sample case?
- Broadly, what is the equivalence?
- Extensions to ANOVA and Chi Square TOI

Two –sample environment	Test statistic	CIs
z-test	$z=\frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$(\bar{x}_1-\bar{x}_2)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
t-test (independent, equal variance)	$t=\frac{\bar{X}_1-\bar{X}_2}{s_p\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$(\bar{x}_1-\bar{x}_2)\pm t\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
t-test (independent, unequal variance)	$t=\frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$(\bar{x}_1-\bar{x}_2)\pm t_{\frac{\alpha}{2},n_1+n_2-2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
Paired t-test	$t=\frac{\bar{di}-d_0}{(\frac{S_{di}}{\sqrt{n}})}$	$\bar{di}\pm t_{\alpha/2,n-1}\frac{S_{di}}{\sqrt{n}}$
f-test	$F=\frac{s_1^2}{s_2^2}$	$\frac{s_1^2/s_2^2}{F_{1-(\alpha/2)}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{s_1^2/s_2^2}{F_{\alpha/2}}$
z-proportion test	$z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\hat{p}_1-\hat{p}_2\pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}$