SJSU SAN JOSÉ STATE UNIVERSITY

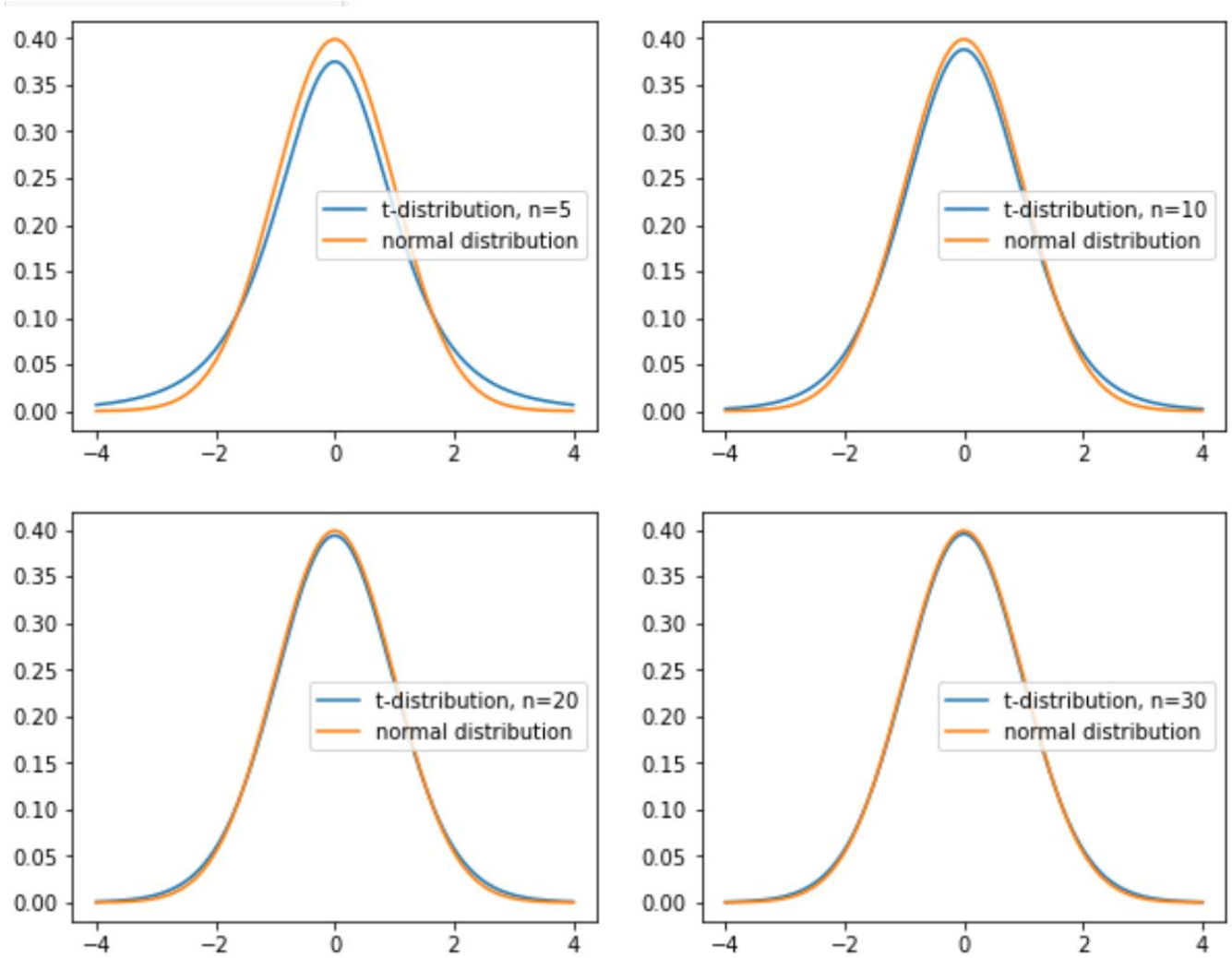
DATA 220: Mathematical Methods for Data Analytics

Dr. Mohammad Masum

Student's t distributions

- Students' t distribution
  - Similar to normal distribution i.e., bell shaped and symmetric
  - However, the tails are heavier than the normal distribution, i.e., more values in the distribution are located in the tail ends
  - Shape of the t-distribution depends on the sample size
    - As sample size increases, the t-distribution approaches towards normal distribution
  - Used to construct confidence interval and conduct hypothesis tests when sample size is small (typically less than 30) and population standard deviation is unknown

Student's t distributions



Student's t distributions

- Students' t distribution

Because the sample size is small, we need to use the t distribution.
For 95% confidence and df = n-1 = 9, t = 2,262.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_confidence_intervals/bs704_confidence_intervals_print.html

Statistical Inference

- Statistical inference is the process of using sample data to make conclusions about a larger population
- A population parameter is a fixed value that describes a characteristic of the population, such as the population mean or standard deviation
- A sample statistic is a value calculated from the sample data that is used to estimate the population parameter
- For example, if we want to estimate the mean height of all adults in a city, we can collect a random sample of adults and calculate the sample mean height
  - We can use this sample mean height to make an inference about the population mean height

Statistical Inference

- There are primarily two types of statistical inference: parameter estimation and hypothesis testing
  - Parameter estimation involves using a sample statistic to estimate a population parameter, along with a measure of how accurate that estimate is (e.g., point estimate, confidence interval)
  - Hypothesis testing involves- making a decision about a population parameter based on a sample statistic, by testing a null hypothesis against an alternative hypothesis

Statistical Inference

Parameter Estimation

	Continuous Variable	Dichotomous Variable (Binary Categorical Variable)
One Sample	mean	proportion or rate
Two Independent Samples	difference in means	difference in proportions
Two dependent, Matched Samples	mean difference	

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_confidence_intervals/bs704_confidence_intervals_print.html

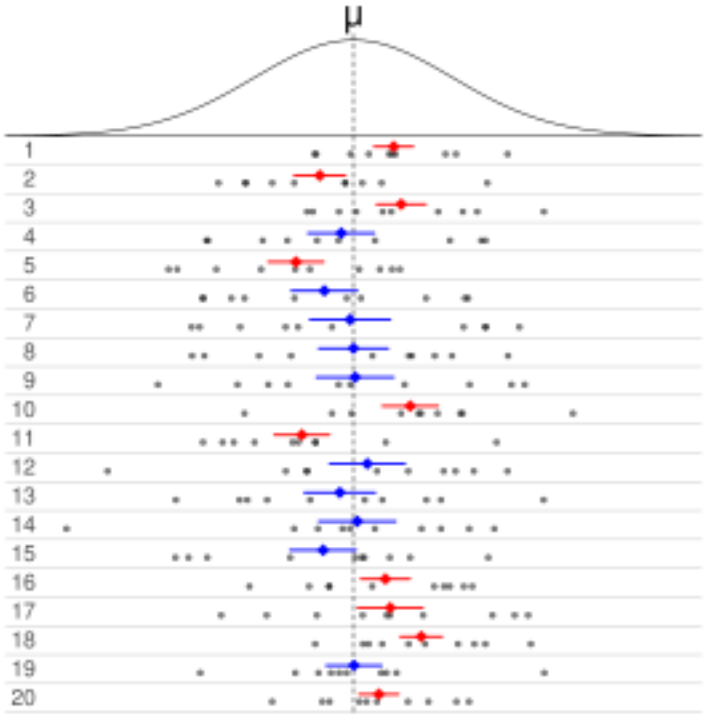
Statistical Inference

- Point estimate-
  - a single numerical value that is used to estimate a population parameter, such as a population mean or proportion
  - estimates are calculated using sample data
  - is subject to sampling error, which is the variation between different samples that would be taken from the same population
  - The accuracy of a point estimate can be assessed by calculating its margin of error, which is a range of values that the true population parameter is likely to fall within

One example of a point estimate is the sample mean, which is used to estimate the population mean. For instance, if we take a random sample of 100 people from a population and find that their average age is 35, we might use this as a point estimate for the population mean age

Statistical Inference

- Confidence interval
  - A range containing the true population parameter with a certain degree of confidence
  - The level of confidence is typically set at 90%, 95% or 99%:
    - Strictly speaking a 95% confidence interval means that if we were to take 100 different samples and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals will contain the true mean value (**μ**)
  - Example: A 95% confidence interval for the population mean height of a certain species of tree might be (10 meters, 15 meters), meaning we are 95% confident that the true population mean height falls between 10 and 15 meters.



Each row of points is a sample from the same normal distribution. The colored lines are 50% confidence intervals for the mean, *μ*. At the center of each interval is the sample mean, marked with a diamond. The blue intervals contain the population mean, and the red ones do not

https://en.wikipedia.org/wiki/Confidence_interval

Statistical Inference

Clin Transl Gastroenterol. 2022 Mar; 13(3): e00463.
Published online 2022 Mar 28. doi: 10.14309/ctg.0000000000000463
PMCID: PMC8963838
PMID: 35142721

Chronic Pancreatitis Is a Risk Factor for Pancreatic Cancer, and Incidence Increases With Duration of Disease: A Systematic Review and Meta-analysis

Sonal Gandhi, MBBS,[1] Jaime de la Fuente, MD,[2] Mohammad Hassan Murad, MD,[3] and Shounak Majumder, MD[2]

▸ Author information ▸ Article notes ▸ Copyright and License information PMC Disclaimer

RESULTS:

Twenty-five cohort and case-control studies met inclusion criteria. Meta-analysis of 12 chronic pancreatitis (CP) studies demonstrated an increased risk of PDAC in patients with CP (SIR: 22.61, 95% confidence interval [CI]: 14.42–35.44). This elevated risk persisted in subgroup analysis of studies that excluded patients diagnosed with PDAC within 2 years of CP diagnosis (SIR: 21.77, 95% CI: 14.43–32.720). The risk was higher in hereditary pancreatitis (SIR: 63.36, 95% CI: 45.39–88.46). The cumulative incidence rates of PDAC in CP increased with follow-up duration. Limited evidence in acute pancreatitis indicates higher PDAC risk in the subset of patients eventually diagnosed with CP. PDAC seems to be uncommon in patients with autoimmune pancreatitis, with 8 reported cases in 358 patients with autoimmune pancreatitis across 4 studies.

Clin Transl Gastroenterol

https://en.wikipedia.org/wiki/Confidence_interval

$$CI = point\ estimate \pm margin\ of\ error(me)$$

$$me = \text{c.v} * \text{SE}$$

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Statistical Inference

- Confidence interval
  - The width of the confidence interval depends on the level of confidence and the sample size
  - Calculated from sample data - based on the sample mean and sample standard deviation, along with a critical value from the t-distribution or normal distribution, depending on the sample size and whether the population standard deviation is known or unknown

Confidence interval for μ

$\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}}$ ➔ when $n \geq 30$; *use Z table*

$\bar{X} \pm t^* \frac{\sigma}{\sqrt{n}}$ ➔ when $n < 30$; *use t table*

Table - Z-Scores for Commonly Used Confidence Intervals

Desired Confidence Interval	Z Score
90%	1.645
95%	1.96
99%	2.576

Statistical Inference

- Practice: An GRE test was given to a simple random sample of 50 students and the sample mean score for quantitative reasoning part was 156.6. According to the official reports by ETS, this test have a standard deviation of 9.58.
  a. Construct a 90% confidence interval for the mean quantitative reasoning score
     - Give a statement of confidence (Interpretation)
     - Is it likely that the population mean score $\mu < 140$
  b. Construct a 95% confidence interval for the mean quantitative reasoning score
  c. Construct a 99% confidence interval for the mean quantitative reasoning score
  – Compare the level of confidence

Statistical Inference

- Practice: An GRE test was given to a simple random sample of 15 students and the sample mean score for quantitative reasoning part was 156.6 and sample standard deviation of 9.58
  a. Construct a 95% confidence interval for the mean quantitative reasoning score
     - Give a statement of confidence (interpretation)

Statistical Inference

- Confidence interval for binary variable – estimating population proportions
  - Sample proportion, $\hat{p} = \frac{x}{n}$; where x is the number of success and n is the sample size
  - Assumptions (so that we can use the normal approximation):
    - sample must be SRS and the population is at least 20 times as large as the sample
    - Sample size should follow: $np > 10 \; and \; n(1-p) > 10$ where n and p *are sample size and population proportion, respectively.*
    - $\boldsymbol{CI = point\ estimate \pm margin\ of\ error}$
      - $\boldsymbol{me = \text{c.v} * \text{SE}}$
        - » $\boldsymbol{\text{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}}$

Statistical Inference

- Practice - In a sample of 400 American adults, it was found that 45% reported that they have confidence in their physician.
  - Find a 95% confidence interval for the proportion of American adults who have confidence in their physician
  - Is it reasonable to conclude that 55% of the adults have confidence in their physician

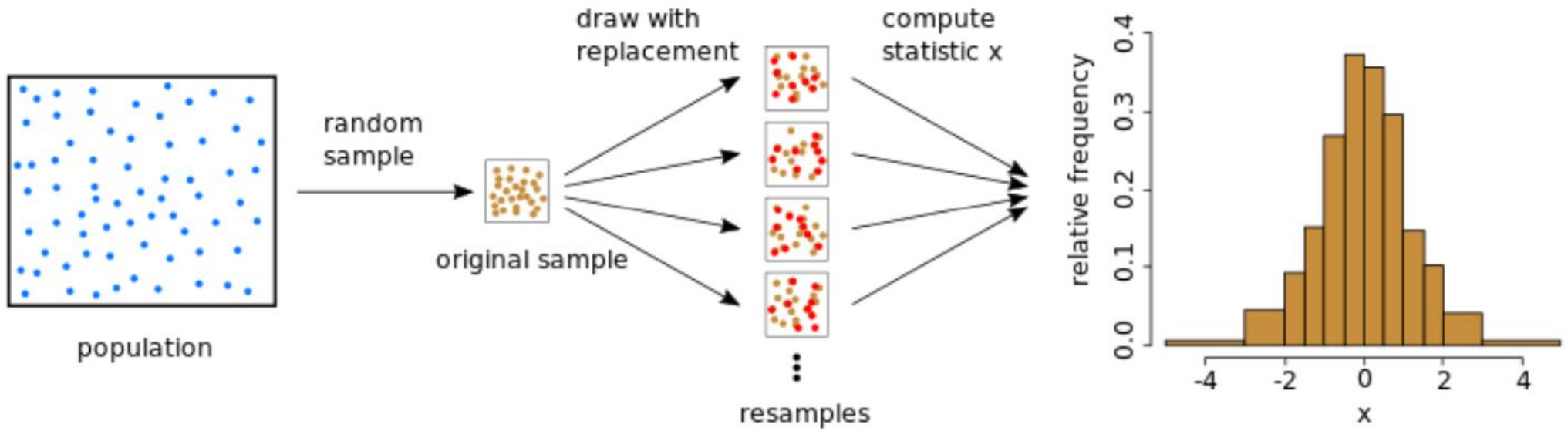
Statistical Inference

- Calculating the sample size

A pollster wants to construct a 95% confidence interval for the proportions of adults who believe that economic conditions are getting better. In the previous poll, proportions was found to be 28%. What sample size is needed to obtain a margin of error of 3%?

Statistical Inference

- Confidence Interval using Bootstrapping (resampling)
  - Repeatedly randomly resampling (with replacement) from your dataset
  - Nonparametric method
  - Can be applied for any kind of parameter estimation (e.g., mean, variance, median, etc.)
  - Limitations: Different (non-significant in general) result each time and Time-consuming for large scale data



https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

Statistical Inference

Procedure to find the bootstrap confidence interval for the mean

1. Draw N samples (N will be in the hundreds, and if the software allows, in the thousands) from the original sample with replacement.
2. For each of the samples, find the sample mean.
3. Arrange these sample means in order of magnitude.
4. To obtain, say, a 95% confidence interval, we will find the middle 95% of the sample means. For this, find the means at the 2.5% and 97.5% percentiles. The 2.5th percentile will be at the position $(0.025)(N+1)$, and the 97.5th percentile will be at the position $(0.975)(N+1)$. If any of these numbers are not integers, round to the nearest integer. The values of these positions are the lower and upper limits of the 95% bootstrap interval for the true mean.

https://www.sciencedirect.com/topics/mathematics/bootstrap-confidence-interval

Statistical Inference

- What is Degrees of Freedom (DF)
  - degrees of freedom refer to the number of independent observations in a sample that can vary without affecting the overall sample statistic of interest
  - Degrees of freedom are often used in calculations involving the t-distribution
    - For example, if you want to calculate a confidence interval for the mean of a sample, you need to use a t-distribution with n-1 degrees of freedom, where n is the sample size
  - Example: imagine you have a classroom with 20 students. If you want to calculate the average height of the students, you are using 20 values. But once you know the average height, you can figure out what the height of the 20th student must be, because the total height of the class is fixed. That means there are only 19 values that are free to vary - so the calculation has 19 degrees of freedom

Hypothesis Test

- Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data.
- It helps to determine if a claim or hypothesis made about a population is supported by the available evidence or not
- It helps to validate the assumptions made about the data and provides a framework for making decisions based on the available evidence.
- Hypothesis testing is a step-by-step process that involves setting up null and alternative hypotheses, choosing an appropriate statistical test, calculating a test statistic, determining the p-value, and making a decision based on the results.

Hypothesis Test

- Null hypothesis (H_0) – a parameter is equal to a specific value or two parameters are equal
- Alternative hypothesis (H_a) – value of the parameter differs
  - Three possible hypothesis
    - Left-Tailed: parameter is less than the value specified by the H_0
    - Right-Tailed: parameter is higher than the value specified by the H_0
    - Two-tailed: parameter is not equal to the value specified by the H_0
- Examples- state the appropriate Null and Alternative hypothesis with curve for rejection region
  - Is the average salary of data scientists more than $100,000 per year?
  - Is the average age of this class less than 25 years?
  - A company that produces light bulbs claims that the average lifespan of their bulbs is 10,000 hours. A consumer group believes that the average may have changed.

Hypothesis Test

- A hypothesis test is like a trail: assume H_o is true → we look for the evidence → decide whether to reject the H_o
- Hypothesis stating conclusion- we reject the H_o or fail to reject the H_o
  - If H_o is rejected based on the evidence: conclude H_a is true
  - If H_o is not rejected based on the evidence: there is no enough evidence to conclude that the H_a is true
- Example –
  - One student from this class thinks that the average salary of data scientists more than $100,000 per year; so the student performs a hypothesis test and rejects the null hypothesis. State an appropriate conclusion
  - Another student from the class also interested in this study and performs a hypothesis test and the student fails to rejects the null hypothesis. State an appropriate conclusion

Hypothesis Test

- Two ways where a hypothesis test may take a wrong decision
  - Type I error: It occurs when we reject the null hypothesis when it is actually true. In other words, we conclude that there is a significant difference or effect when there is none.
  - Type II error: It occurs when we fail to reject the null hypothesis when it is actually false. In other words, we conclude that there is no significant difference or effect when there is one.

Your decision		H_0 *true*	H_0 *false*
	Reject H_0	*Type I* *False Positive* $(1-\alpha)$	*Correct decision* *True Positive* $(1-\beta)$
	Fail to reject H_0	*Correct decision* *True Negative* (α)	*Type II* *False Negative* (β)

Hypothesis Test

- Two ways where a hypothesis test may take a wrong decision
  - Type I error
  - Type II error

Hypothesis Test

- The dean of a business school wants to determine whether the mean starting salary of graduates of her school is greater than $70,000.
  - State the appropriate null and alternative hypothesis
  - Suppose that the true mean is $\mu = \$75{,}000$, and the dean rejects H_0. State an appropriate conclusion. Is this a Type I error, Type II error, or a correct decision?
  - Suppose that the true mean is $\mu = \$65{,}000$, and the dean does not reject H_0. Is this a Type I error, Type II error, or a correct decision?

Example of Type I & Type II Error

- Determine whether the conclusion is a type I error, type II error, or correct decision
  - A test is made with H0: mu = 40 versus H1: mu > 40. The true value of mu is 45 and H0 is not rejected.
  - H0 is False ➔ Type II Error
  - A test is made with H0: mu = 1.5 versus H1: mu > 1.5. The true value of mu is 1.5 and H0 is not rejected.
  - H0 is true ➔ correct decision
  - A test is made with H0: mu = 9 versus H1: mu < > 9. The true value of mu is 9 and H0 is rejected.
  - H0 is true ➔ Type I error
  - A test is made with H0: mu = 15 versus H1: mu < 15. The true value of mu is 11 and H0 is rejected.
  - H0 is false ➔ correct decision

Hypothesis Test

- A statistical test is a way to determine if a sample result is unlikely to have occurred by chance
- P-values
  - The reasoning of a statistical test is that the probability of getting that sample result is unusual(?), then the null hypothesis must no longer be true
    - The probability we calculate is the P-value
  - a measure of how much evidence we have against the null hypothesis
  - the smaller the p-value, the stronger the evidence against the null hypothesis
- Use cases
  - A/B testing, p-values are used to determine whether the results of a test are statistically significant and whether a change in a website or application should be implemented
  - In regression analysis, p-values are used to determine the significance of the coefficients and the overall model

Hypothesis Test

- How we can tell the obtained p-value is unusual?
  - if it is smaller than the predetermined level of significance (alpha, α)
  - the level of significance represents the probability threshold for considering a result as statistically significant
  - Common level of significance used in hypothesis testing: 0.05, 0.01, 0.1

Hypothesis Test

- Hypothesis testing – p-Value Approach
  - Comparing p-value to significance level
  - Steps of conducting hypothesis using C.V.
    - State null and alternative hypothesis
    - Assume the null hypothesis is true and calculate the test statistics using the sample data information
    - Find p-value from the test statistics
    - Compare the p-value with the given significance level
      - If p-value $\leq \alpha$ ➔ we reject the H_0
      - If p-value $> \alpha$ ➔ we fail to reject the H_0

Hypothesis Test

- Hypothesis testing – Critical Value Approach
  - Comparing the test statistics (z-test or t-test) to some cutoff value, called critical value
    - If test statistics is more extreme (higher or lower) that the critical value ➔ H_o is rejected in favor of alternative
    - If test statistics is not extreme ➔ fail to reject H_o
  - Steps of conducting hypothesis using C.V.
    - State null and alternative hypothesis
    - Assume the null hypothesis is true and calculate the test statistics using the sample data information
    - Find the critical value from the given significance level using a table (t- or normal) or using calculator or python
    - Compare test statistics and the critical value

Hypothesis Test

Standard Normal Probabilities

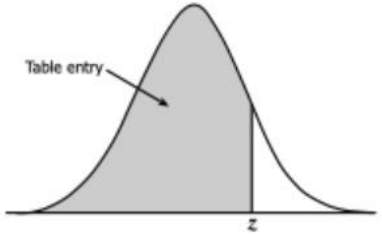


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

SJSU SAN JOSÉ STATE UNIVERSITY

Hypothesis Test

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
					Confidence Level						

Hypothesis Test

- Test statistics
  - a test statistic is a numerical value that is calculated from the sample data and is used to test a hypothesis
  - It measures the difference between the sample statistic and the hypothesized population parameter, taking into account the variability of the sample.

Z-statistics for mean

$$Z^* = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\frac{\sigma}{\sqrt{n}}}$$

Z-statistics for proportion

$$Z^* = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

t-statistics for mean

$$t^* = \frac{\bar{X} - \mu_{\bar{X}}}{\frac{s}{\sqrt{n}}}$$

Hypothesis Test

- If the population proportion (p) is known, then use the sample proportion standard deviation formula: $\sqrt{\frac{p(1-p)}{n}}$
- If you only know the sample proportion $\hat{p}$, then use the standard error of the sample proportion formula: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Z-statistics for proportion

$$Z^* = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Hypothesis Test

- Z tests or T-tests???
  - Hypothesis testing for mean when-
    - We know the population standard deviation
    - We do not know the population standard deviation
- Practice:
  - It is reported that the average price of a gallon of gas in your state is $5.816 with a population standard deviation of $0.027. You believe that in your county the price is different. You take a random sample of prices from 15 gas stations, recording a sample mean cost of $5.847. We are assuming that gas prices are normally distributed. Conduct a hypothesis test to test the claim that the gas prices are different in your county at the .05 significance level.

Hypothesis Test

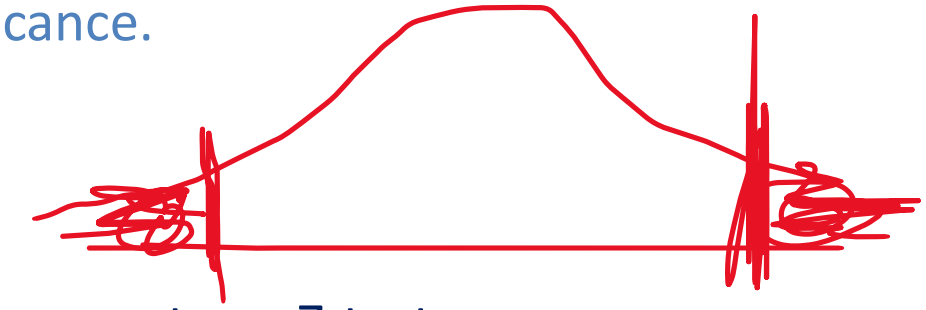
- A study was conducted to determine whether a standard clerical test would need revision for use on a different style of keyboard. The test has had a mean of 243.5. In a sample of 22 subjects, the sample mean on the new keyboard is 237.2 with a sample standard deviation of 35.3. Test the claim that mean for the test on the new keyboard is different than the test on the previous keyboard. Assume the scores in the population are normally distributed.

Hypothesis Test

- Practice: It is reported that the average college student sleeps 5.2 hours per night during the semester. A local college is concerned that the average number of hours may be lower than that for their students. They conduct a study of 75 students and find a mean sleep time of 4.9 hours with a sample standard deviation of 1.3 hours. Conduct a hypothesis test at a .05 significance level to determine whether the mean number of hours of sleep for college students at this college is significantly lower.

Hypothesis Test

- Practice: A random sample of 100 pairs of ladies shoes had a mean size of 8.3. Assume the population standard deviation is 1.5. Can you conclude that the mean size of ladies shoes differs from 8? Use the 0.01 level of significance.



1. $H_0: \mu = 8 \; vs. \; H_a: \mu \neq 8$
2. $\alpha = 0.01; n = 100 \; ; \bar{x} = 8.3; \sigma = 1.5$
3. Since population standard deviation is known, we must use Z-test
4. $Z^* = \frac{8.3-8}{\frac{1.5}{\sqrt{(100)}}} = 2$ ➔ Now, we will look for the region which is higher than 2; that is, we must find $P(Z^* > 2)$; from the z-table we can find the value, $P(Z^* < 2) = 0.9722$; therefore, $P(Z^* > 2)$ = 1- 0.9722 = 0.0228; since this is a two-tailed test, we must multiply the p-value by 2; therefore, (the total area under the rejected region) p-value = 2*0.0228 = 0.0456
5. Since the p-value > alpha (0.01) ➔ we fail to reject the H_0 and conclude that there is no enough evidence that the mean size of ladies/ shoes differs from 8.

z	.00
0.0	.5000
0.1	.5398
0.2	.5793
0.3	.6179
0.4	.6554
0.5	.6915
0.6	.7257
0.7	.7580
0.8	.7881
0.9	.8159
1.0	.8413
1.1	.8643
1.2	.8849
1.3	.9032
1.4	.9192
1.5	.9332
1.6	.9452
1.7	.9554
1.8	.9641
1.9	.9713
2.0	.9772

Hypothesis Test

- Hypothesis test for proportions
  - What are the assumptions?
- Practice: A leading research firm reports that 33% of young adults have completed a college degree. One state leader who is trying to create more scholarship money for college believes that the rate is lower than that in his state. The state leader commissions a study that surveys 500 young adults in the same age bracket as the national study to see if they have completed a college degree. Of the 500 interviewed, 152 have completed a college degree. Conduct a hypothesis test to see if the state leader has evidence to support his claim.

Hypothesis Test

A random sample of 1500 U.S. adults is taken. They are asked whether they approve or disapprove of the current president's performance so far (i.e., an approval rating). Of the 1500 surveyed, 660 respond with "approve".

Suppose we want to test if the proportion is different than 40%. In other words, we want to test the following hypotheses at significance level 5%.

Hypothesis Test

- Using data to conduct hypothesis tests: Assumptions-
  - **Random sampling**: The sample should be randomly selected from the population of interest
  - **Independence**: The sample observations should be independent of each other
  - **Normality**: Sample drawn from a population that is approximately normal or the data should follow a approximately normal distribution or the sample size should be large enough to apply the central limit theorem
    - Check outliers, skewness, and multimodality

Hypothesis Test

- Using data to conduct Suppose we want to test the hypothesis that the mean weight of apples produced by a local farmer is 150 grams. We randomly select 6 apples and weigh them, getting the following weights (in grams): 155, 148, 153, 152, 147, and 151.
  - Use a dot plot to explain whether it is reasonable to assume the conditions for performing a hypothesis test are satisfied?
  - Can we reject the null hypothesis at the 5% significance level?

Hypothesis Test – two sample t-test

- Two sample t-test is an extension of one sample t-test
  - two independent groups of data have different means
  - It compares the means of two samples to see if they are statistically significantly different
  - It assumes that the data in each group are normally distributed and have equal variances

  - Null hypothesis: $H0: \mu1 - \mu2 = 0$
  - Alternative hypothesis: $Ha: \mu1 - \mu2 \neq 0$

Hypothesis Test – two sample t-test

- Paired test (dependent t test / repeated measures t test): each measurement in one sample is matched or paired with a particular measurement in the other sample
    - Used to compare the means of two dependent samples

$$t^* = \frac{\overline{x_d}}{\frac{s_d}{\sqrt{n}}}$$

Hypothesis Test – two sample t-test (Paired)

	Before	After	Difference
0	22	26	4
1	27	29	2
2	18	22	4
3	20	24	4
4	25	27	2
5	21	23	2
6	19	21	2
7	24	25	1
8	23	24	1
9	26	28	2

```
14 # Conduct paired t-test
15 alpha = 0.05
16 t_statistic, p_value = ttest_rel(df['Before'], df['After'])
17
18 # Determine if null hypothesis is rejected
19 if p_value < alpha:
20     print(f"p-value: {p_value}, reject null hypothesis")
21 else:
22     print(f"p-value: {p_value}, fail to reject null hypothesis")
23
```

```
p-value: 0.00011598239506288895, reject null hypothesis
```

Hypothesis Test – two sample t-test

- Unpaired test (independent sample): samples taken from one population have no relationship with samples taken from other population
    - Used to compare the means of two independent samples

$$t = \frac{\bar{x}_1 - \bar{x_2} - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hypothesis Test – two sample t-test (equal variance)

- Suppose we want to test the hypothesis that the mean weight of apples produced by Farm A is different from the mean weight of apples produced by Farm B. We randomly select 10 apples from each farm and weigh them, getting the following weights (in grams):
  - Farm A: 135, 142, 149, 146, 138, 152, 147, 141, 148, 136
  - Farm B: 150, 142, 154, 143, 152, 145, 148, 146, 149, 147
- Assume equal variance of the two samples

```
 8 # calculate t and p-value
 9 t, p = ttest_ind(farm_A, farm_B)
10
11 # print results
12 print("t =", t)
13 print("p =", p)
14
15 if p < 0.05:
16     print("We reject the null hypothesis that the means are equal.")
17 else:
18     print("We fail to reject the null hypothesis that the means are equal.")
19
```

```
t = -1.9021170377133751
p = 0.07328147862748967
We fail to reject the null hypothesis that the means are equal.
```

Hypothesis Test – two sample t-test (Unequal Variance)

- A hospital wants to determine whether there is a difference in the recovery times between two different surgical procedures. They randomly select 10 patients who have undergone procedure A and 8 patients who have undergone procedure B. The results are as follows:
  - Procedure A: 10.2, 9.7, 12.1, 11.9, 11.5, 10.8, 11.3, 12.5, 10.1, 9.9
  - Procedure B: 8.7, 7.8, 9.3, 8.5, 10.1, 9.5, 10.4, 7.9
  - Assuming the variances are not equal, is there a significant difference in the recovery times between the two procedures?

```
1 import scipy.stats as stats
2
3 procedureA = [10.2, 9.7, 12.1, 11.9, 11.5, 10.8, 11.3, 12.5, 10.1, 9.9]
4 procedureB = [8.7, 7.8, 9.3, 8.5, 10.1, 9.5, 10.4, 7.9]
5
```

```
1 t_stat, p_val = stats.ttest_ind(procedureA, procedureB, equal_var=False)
2 print('t-statistic:', t_stat)
3 print('p-value:', p_val)
4
```

```
t-statistic: 4.249699062856683
p-value: 0.000663519953237813
```