An overview of inferential statistics

Amm Quamruzzaman
ISF, UC Berkeley

March 8, 2022

Outline

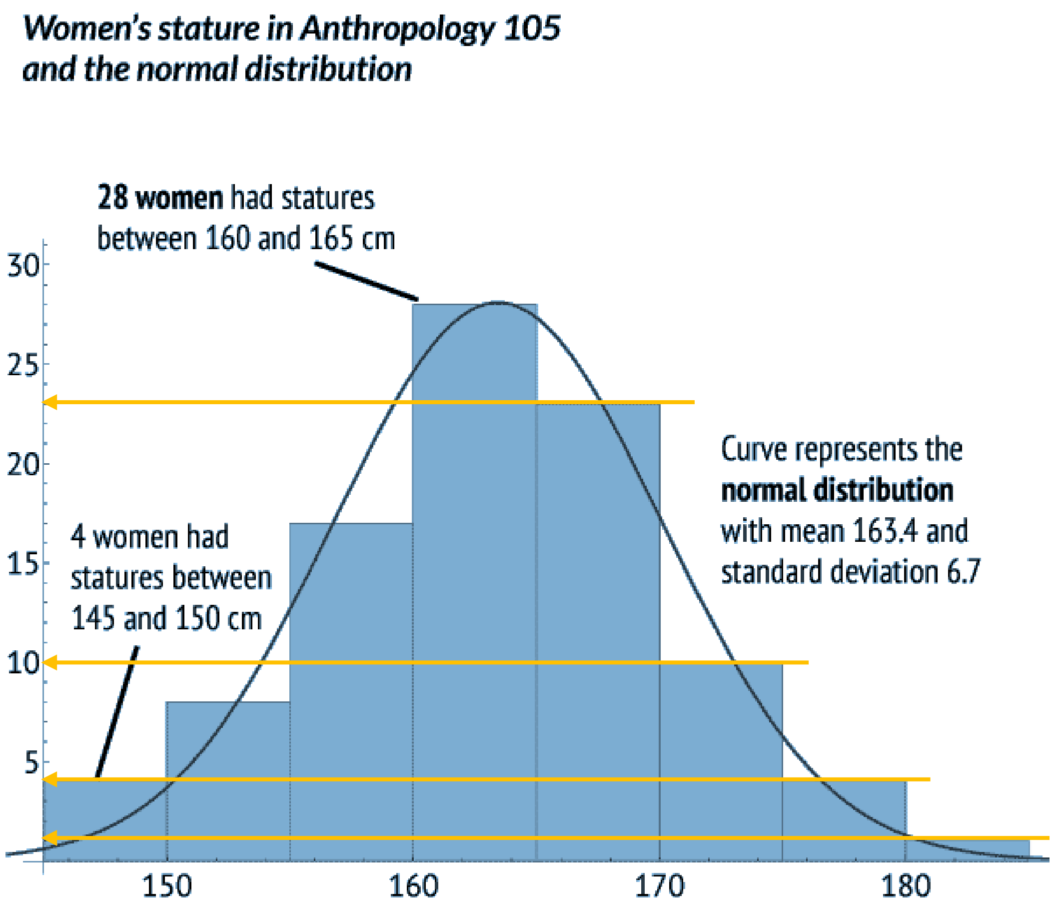
I. Introduction
II. The theory of probability
III. Probability and hypothesis testing
IV. The *z*-distribution and hypothesis testing
V. The *t*-distribution and hypothesis testing
VI. The $\chi 2$-distribution and hypothesis testing
VII. ANOVA for hypothesis testing

I. Introduction

- Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability.
- We use a sample to make inferences about the population.
- Of course, from a sample, we cannot make exact inferences.
- There will be some uncertainty; this uncertainty can be described precisely using the theory of probability.
- Probabilities are proportions, and so can be multiplied by 100 and expressed as percentages: 0% = impossible to occur, 50% = as likely to occur as not, 100% = certain to occur.
- Alternatively, probability values vary from 0 (impossible) to 1 (certain). The values cannot be negative.

II. The theory of probability

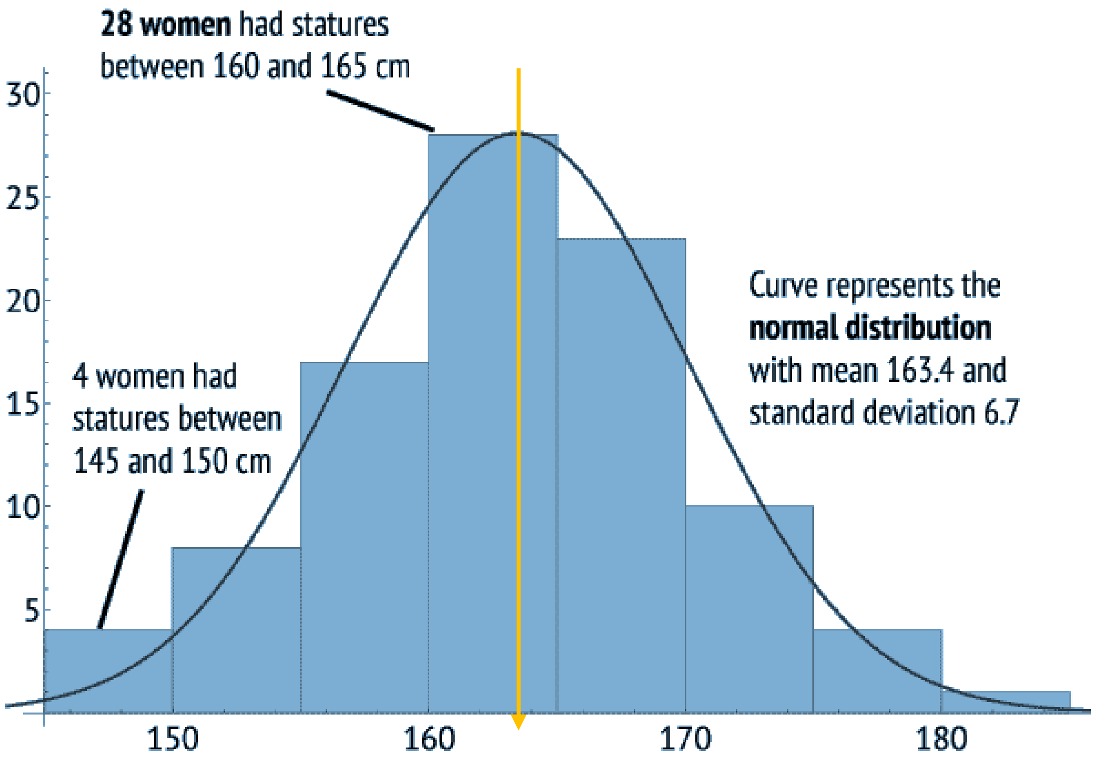
- Based on coin tossing/dice rolling.
- The probability of an outcome is the observed frequency of that outcome divided by the total number of observed outcomes.
- In the example (right), the probability of women (n=95) having statures between 160 and 165 cm is 28/95 = 0.29.
- What is the probability of women having statures over 165 cm?
- (23+10+4+1)/95 = 0.40.



II. The theory of probability...

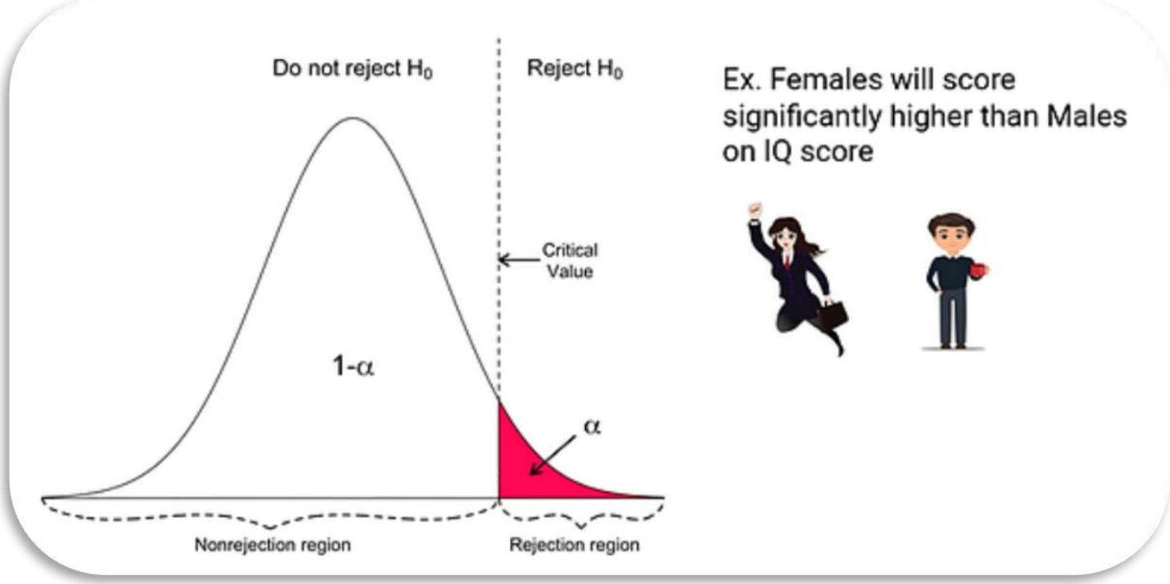
- The probability distribution is a curve describing an idealized frequency distribution of a variable from which it is possible to get the probability with which specific values of that variable will occur.
- Certain shapes of distribution are quite common; they are called the normal distribution. For these shapes mathematicians have worked out the probability scores or *p* values.

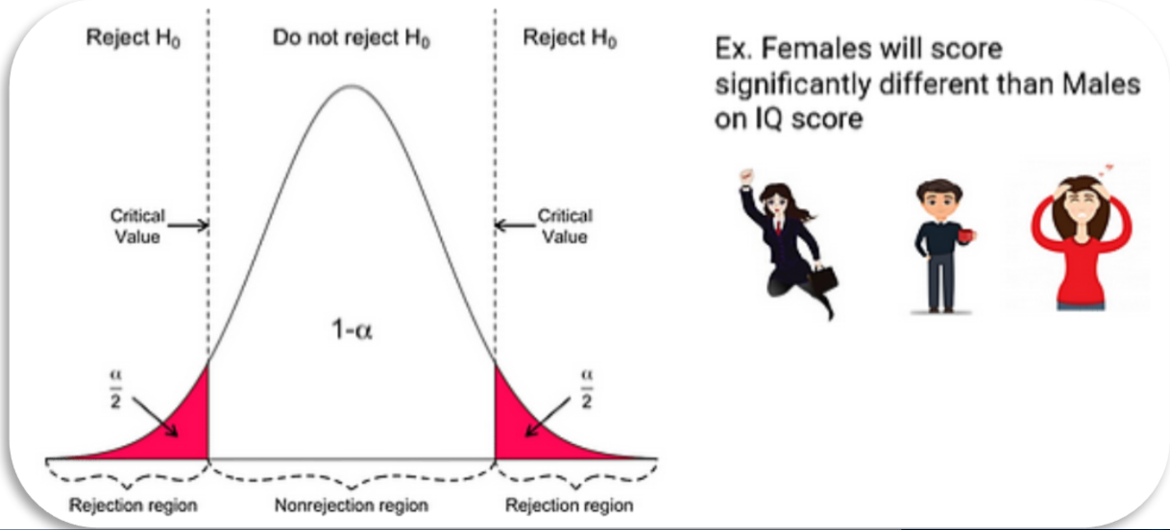
Women's stature in Anthropology 105 and the normal distribution



III. Probability and hypothesis testing

- The *p* value is the probability that the samples are from the same population for an outcome (DV).
- Usually, the hypothesis we are testing (called null hypothesis) is that the samples (groups) differ on the outcome.
- The *p* value determines whether or not we reject the null hypothesis and accept the alternative hypothesis.



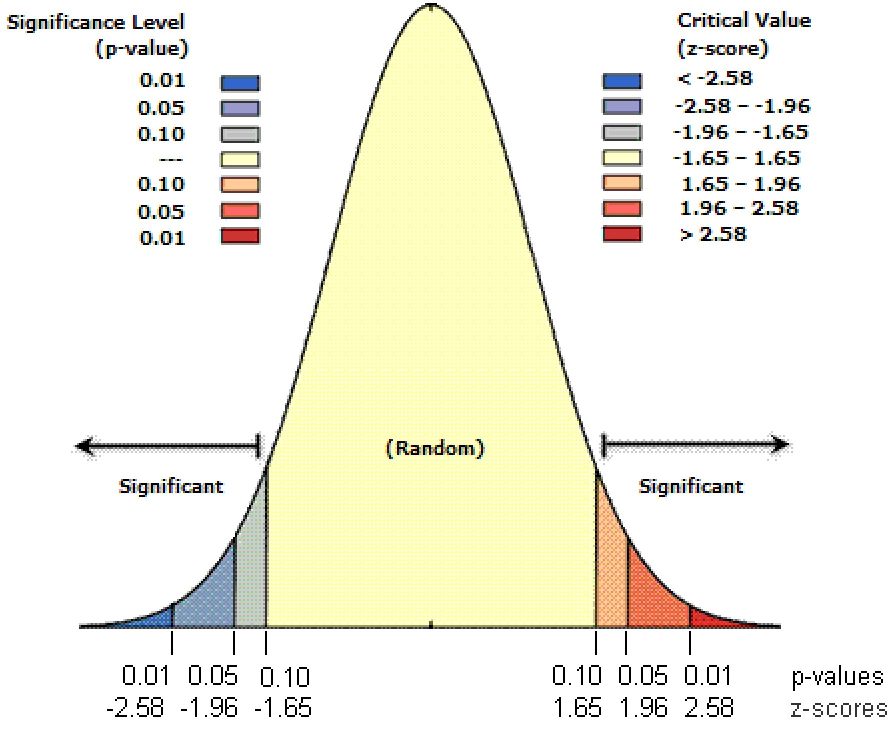


III. Probability and hypothesis testing…

- Conventional *p* values are: $^{***}p \leq 0.001$; $^{**}p \leq 0.01$; $^{*}p \leq 0.05$ and are referred to as the levels of significance.
- If the *p* value is large (> 0.05), accept the null hypothesis (i.e., the alternative hypothesis is not true).
- If the *p* value is small (≤ 0.05), reject the null hypothesis.
- There is always a possibility that we are making a mistake in rejecting the null hypothesis.
- This is called a Type I Error – rejecting the null hypothesis when it is true. If this probability is ≤ 0.05, we are still 95% confident but not 100% safe from committing an error.
- A Type II Error – accepting the null hypothesis when it is false.

III. Probability and hypothesis testing…

- Different hypothesis tests use different test statistics based on the shapes of the probability curve.
- There are, for example: *z* distribution, *t* distribution, chi-square (χ2) distribution, and *F*-distribution.
- They are all defined by an equation that enables us to calculate test statistics. We use the test statistics to decide about a hypothesis.



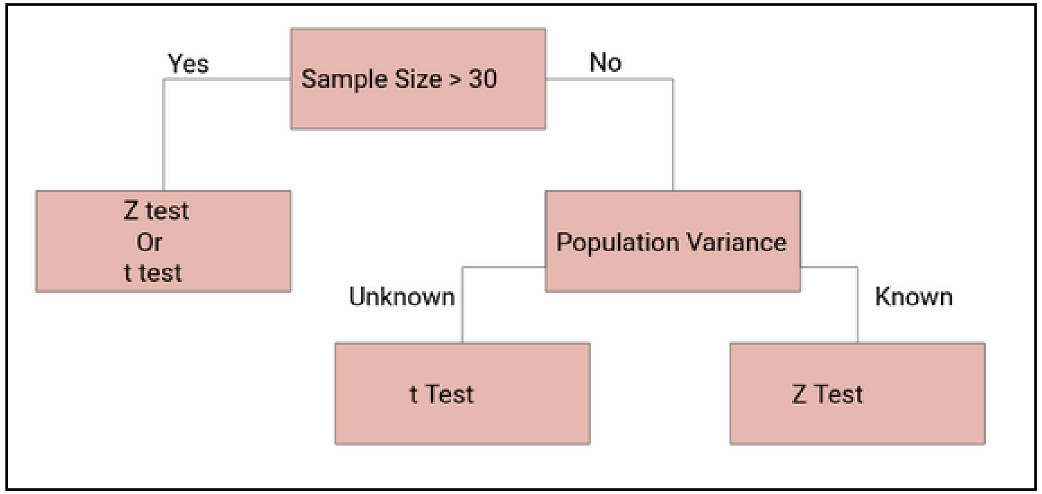
IV. When to use which test?

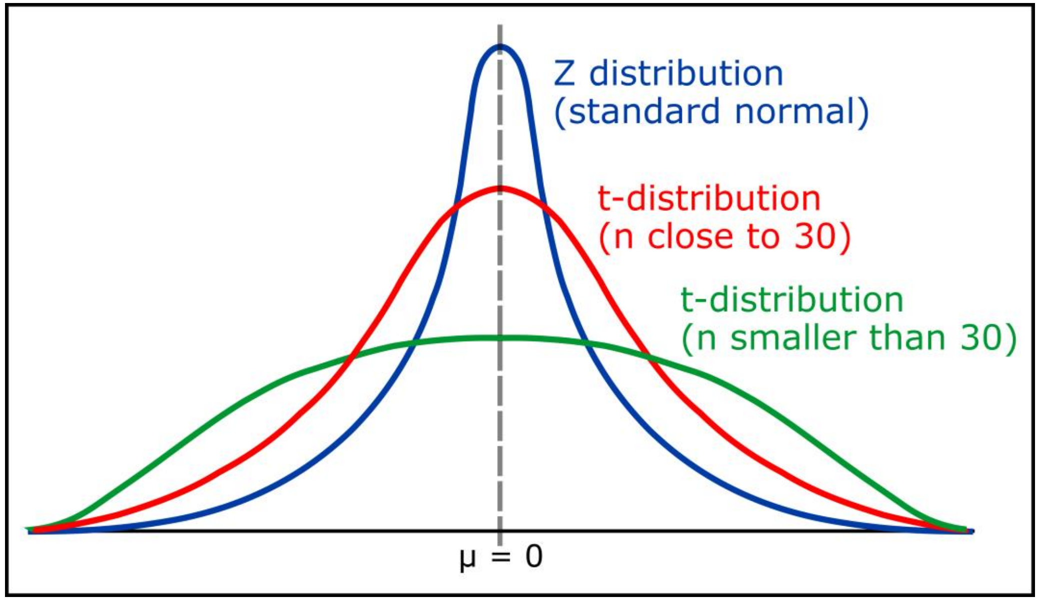
- Depends on sample size, number of groups, types of variables, and the objectives of a test.

Test statistic	Associated test	Sample size	Information given	Distribution	Test question
z-score	z-test	Two populations or large samples (n > 30)	• Standard deviation of the population (this will be given as σ) • Population mean or proportion	Normal	Do these two populations differ?
t-statistic	t-test	Two small samples (n < 30)	• Standard deviation of the sample (this will be given as *s*) • Sample mean	Normal	Do these two samples differ?
f-statistic	ANOVA Regression	Three or more samples	• Group sizes • Group means • Group standard deviations	Normal	Do any of these three or more samples differ from each other?
chi-squared	chi-squared test	Two samples	• Number of observations for each categorical variable	Any	Are these two categorical variables independent?

V. The *z*-distribution and hypothesis testing

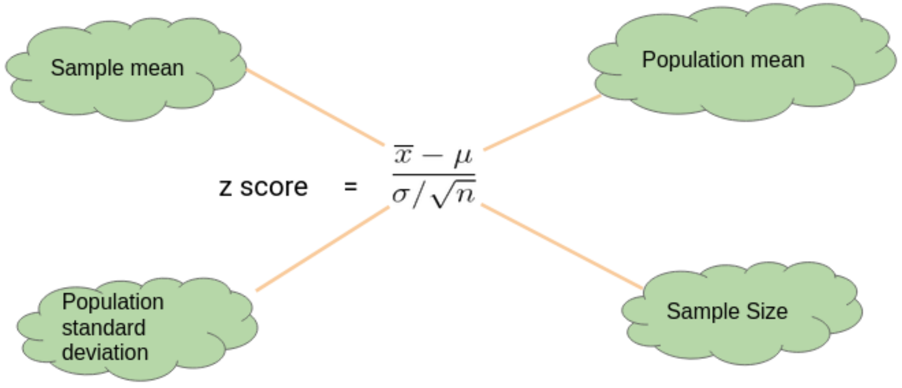
- *Are rich countries suffering more from Covid-19 than poor countries?*
- A z-test is a statistical way of testing a hypothesis when
  - We know the population variance, or
  - We do not know the population variance, but our sample size is larger than 30.
- If we have a sample size of less than 30 and do not know the population variance, we use a *t*-test.





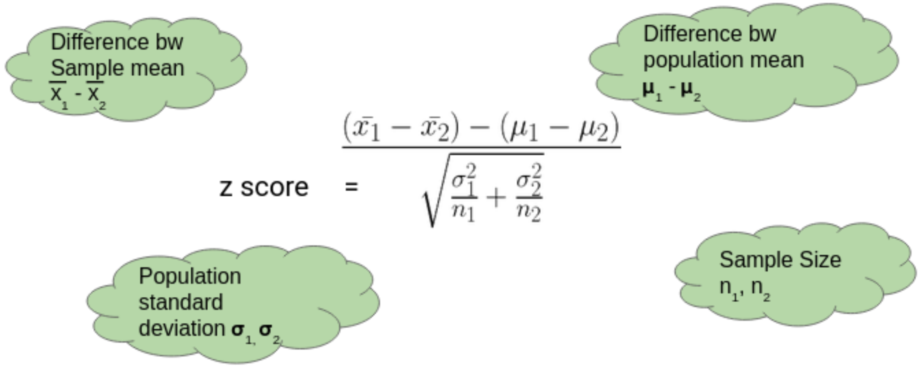
V. The *z*-distribution…

- One-sample *z*-test:



$$\text{z score} = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

- Two-sample *z*-test:



$$\text{z score} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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
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										

https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/

VI. The *t*-distribution and hypothesis testing

- The *t*-distribution is like the normal distribution with its bell shape but has heavier tails (a greater chance for extreme values).
- William Sealy Gosset (1905) first published a *t*-distribution under the name Student. (Erich Lehman described *z*-test in 1986.)
- Student's *t*-test is used to determine if there is a significant difference between the means of two groups in terms of a continuous variable.
- It is useful when the sample size is small (n<30), and the population standard deviation is unknown.
- Example: After tutoring only boys in statistics, is there a difference in the test scores between boys and girls?

VI. The *t*-distribution...

Observation #	Boys' scores	Girls' scores
1	8	6
2	7	7
3	10	8
4	10	7
5	9	7
6	8	10
7	7	8
8	8	7
9	10	8
10	9	6

- Formula: $t = \dfrac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

	Boys' scores	*Girls' scores*
Mean	8.6	7.4
Variance	1.37777777	1.37777777
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	2.286002286	
P(T<=t) one-tail	0.017296716	
t Critical one-tail	1.734063607	
P(T<=t) two-tail	0.034593432	
t Critical two-tail	2.10092204	

VII. The χ2 distribution and hypothesis testing

- *Z*-test and *t*-test cannot be used for categorical variables. For example, if you want to test the following hypothesis:
  - *In a relationship, happiness depends on whether a person is faithful or unfaithful to the other.*
- Say, to test this hypothesis you interviewed 506 men aged 18-50 years.
- The DV is "relationship happiness" (whether a man was happy or unhappy) and the IV is "infidelity" (whether a man was faithful or unfaithful)

VII. The χ2 distribution...

- Say, we get a distribution of our sample of 506 men as follows:

Table 1 *Contingency table showing how many men engaged in infidelity or not, based on how happy they were in their relationship. Data from Table 3 of Mark et al.*[92]

		Infidelity		
		Unfaithful	Faithful	**Total**
Happiness in relationship	Unhappy (greater than the median)	56	101	**157**
	Happy (median or less)	62	287	**349**
	Total	**118**	**388**	**506**

- Hypothesis tests can be performed on contingency tables using χ2 statistic.

VII. The χ2 distribution...

- To calculate χ2 statistic, we need to calculate expected frequencies for each cell / each combination of categories:

$$\text{model}_{\text{Unhappy,Unfaithful}} = \frac{\text{RT}_{\text{Unhappy}} \times \text{CT}_{\text{Unfaithful}}}{n} = \frac{157 \times 118}{506} = 36.61$$

$$\text{model}_{\text{Unhappy, Faithful}} = \frac{\text{RT}_{\text{Unhappy}} \times \text{CT}_{\text{Faithful}}}{n} = \frac{157 \times 388}{506} = 120.39$$

$$\text{model}_{\text{Happy,Unfaithful}} = \frac{\text{RT}_{\text{Happy}} \times \text{CT}_{\text{Unfaithful}}}{n} = \frac{349 \times 118}{506} = 81.39$$

$$\text{model}_{\text{Happy,Faithful}} = \frac{\text{RT}_{\text{Happy}} \times \text{CT}_{\text{Faithful}}}{n} = \frac{349 \times 388}{506} = 267.61$$

- Next step: compare these values to those that we actually observed – calculate the differences between observed (*O*) and expected (*E*) frequencies.

VII. The χ2 distribution...

		Infidelity		
		Unfaithful	Faithful	**Total**
Happiness in relationship	Unhappy (greater than the median)	36.61	120.39	**157**
	Happy (median or less)	81.39	267.61	**349**
	Total	**118**	**388**	**506**

- Use this formula to calculate the χ2 statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(56-36.61)^2}{36.61} + \frac{(101-120.39)^2}{120.39} + \frac{(62-81.39)^2}{81.39} + \frac{(287-267.61)^2}{267.61}$$

$$= \frac{19.39^2}{36.61} + \frac{-19.39^2}{120.39} + \frac{-19.39^2}{81.39} + \frac{19.39^2}{267.61}$$

$$= 10.27 + 3.12 + 4.62 + 1.40$$

$$= 19.41$$

VII. The χ2 distribution...

- Now, find out the *p*-value associated with the χ2 statistic from the chi-square distribution table.
- The shape of chi-square distribution is affected by the degrees of freedom, which are calculated as (*r*-1)(*c*-1) in which *r* is the number of rows and *c* is the number of columns.
- In this case, *df* = (2-1)(2-1) = 1.
- Since the test statistic (19.41) is much bigger than either critical value, we conclude that the hypothesis is accepted (i.e., there is a statistically significant relationship b/ being faithful in a relationship and being happy).

	p	
df	0.05	0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48

Critical values for *df* = 1

Critical values, *p* = 0.05

Critical values, *p* = 0.01

VIII. ANOVA for hypothesis testing

- ANOVA is used to determine whether there are any statistically significant differences between the means of three or more unrelated groups. Ronald Fisher invented ANOVA in 1918.
- The Formula for ANOVA is:
  - F=MSE/MST, where:
    - F=ANOVA coefficient
    - MST=Mean sum of squares due to treatment
    - MSE=Mean sum of squares due to error
- The ANOVA test results are used in an *F*-test to generate additional information that aligns with regression models.