CDS - STATISTICS CHEATSHEET

TEST SIZES

Population: The entire set of possible observations in which we are interested in

Sample: A subset of the population from which the information is actually collected
Gathering the complete population's data is not always possible due to barriers such as time, accessibility, or cost. So, we often gather information from a smaller subset of the population, known as a sample.

VARIABLES

Variables are properties of some event, object or person that can take on different values.

Discrete variables can take only certain values
Example: a household could have 2 children or 4 children, but not 2.53 children

Continuous variables can take any value withing the range of the scale
Example: the temperature recordings during the day, height measurements of hospital patients

MEASURES OF CENTRE

Mean is the sum of all the values in the sample divided by the number of values in the sample/population. μ is the mean of the population; $\bar{x}$ is the mean of the sample

Median is the value separating the higher half of a sample/population from the lower half. It is found by arranging all the values in ascending or descending order and taking the middle one (or the mean of the middle two if there are even number of values)

MEASURES OF SPREAD:

Variance: Measures dispersion around the mean; determined by averaging the squared differences of all the values from the mean
Variance of a population is σ^2: $\sigma^2 = \frac{\sum(x-\mu)^2}{n}$
It can be calculated by subtracting the square of the mean from the average of the squared scores: $\sigma^2 = \frac{\sum x^2}{n} - \mu^2$
Variance of a sample is S^2. $S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ (note the $n-1$)
It can be calculated by: $S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$

Standard deviation: Square root of the variance
Also measures dispersion around the mean but in the same units as the values (instead of square units with variance)
σ is the standard deviation of the population and S is the standard deviation of the sample

Standard error: An estimate of the standard deviation s
of the sampling distribution – the set of all samples of size n that can be taken from a population. It reflects the extent to which a statistic changes from sample to sample
For a mean, $\frac{s}{\sqrt{n}}$
For the difference between two means,

- Assuming equal variances $\sqrt{S^2(\frac{1}{n_1}+\frac{1}{n_2})}$
- Unequal variances $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Confidence Interval:

Confidence Intervals describe the variability surrounding the sample point estimate (the wider the interval, the less confident we can be about the estimate of the population mean).

The equation for a 95% Confidence Interval for the population mean when the population standard deviation is unknown and the sample size is large (over 30) is

$$\bar{X} \pm 1.96\frac{s}{\sqrt{n}}$$

$$\bar{X} = sample\ mean, \qquad n = number\ in\ sample, \qquad \frac{s}{\sqrt{n}} = standard\ error$$

For example, if sample A of 30 babies born in 2015 had a mean weight of 7.19 lbs with a standard deviation of 1.4, the 95% Confidence Interval for the population mean of all babies is: 7.19 ± 1.96 · 1.4/√30 = 7.19 ± 0.5 = (6.7, 7.7) i.e., we would expect the population mean to be between 6.7 lbs and 7.7 lbs. Confidence intervals give a range of values within which we are confident (in terms of probability) that the true value of a population parameter lies. A 95% CI is interpreted as 95% of the time the CI would contain the true value of the population parameter.

HYPOTHESIS TESTING:

Hypothesis testing is an objective method of making decisions or inferences from sample data (evidence). Typically, we compare what we have observed to what we expect if one of the statements (Null Hypothesis) was true.

Null hypothesis (H0) is a statement about the population & sample data used to decide whether to reject that statement or not. Typically, the statement is that there is no difference between groups or no association between variables.

Alternate Hypothesis (H1) The alternative hypothesis (typically) states that observations are the result of a real effect, plus a chance variation - not purely chance. It is often the research question and varies depending on whether the test is one or two tailed.

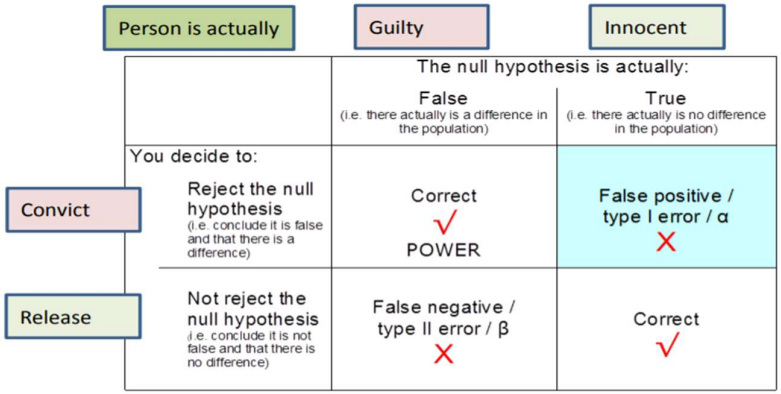
Significance Level (or alpha (α)) is the probability of rejecting the null hypothesis when it is true, (also known as a type 1 error). This is decided by the individual but is normally set at 5% (0.05) which means that there is a 1 in 20 chance of rejecting the null hypothesis when it is true.

Test statistic is a value calculated from a sample to decide whether to accept or reject the null (H0) and varies between tests. The test statistic compares differences between the samples or between observed and expected values when the null hypothesis is true.

p-value is the probability of obtaining a test statistic at least as extreme as what we have obtained, if the null is true and there really is no difference or association in the population of interest. A significant result is when the p-value is less than the chosen level of significance (usually 0.05).

Example (Hypothesis testing): Members of a jury have to decide whether a person is guilty or innocent based on evidence presented to them. **Null**: The person is innocent **Alternate:** The person is guilty. The null can only be rejected if there is enough evidence to disprove it and the jury do not know whether the person is really guilty or innocent so they may make a mistake. If a court case was a hypothesis test, the jury consider the likelihood of innocence given the evidence and if there's less than a 5% chance that the person is innocent, they reject the statement of innocence. In reality, although the person is either

actually Guilty (null false) or Innocent (null true) but we can only conclude that there is evidence to suggest that the null is false or not enough evidence to suggest it is false.



Hypothesis Testing

Formal Question: "If Aspirin had no effect, what is the probability that this result occurred by chance?"

Step 1:

- H_0 Null hypothesis. Aspirin had no effect $p_1 = p_2$
- H_AAlternate hypothesis. Aspirin does reduce heart attack rate $p_1 > p_2$

Step 2:

- Under Null hypothesis $p_1 = p_2 = p$
- Under Null hypothesis

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

- Test Statistic $Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)}$;
- Under H_0 $p_1 = p_2 => p_1 - p_2 = 0$

Step 3:

- $\hat{p} = \frac{378}{22071} = 0.0171,\ 1 - \hat{p} = 0.9829$
- $SE_0(\hat{P}_1 - \hat{P}_2) = 0.00175$
- $Z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{P}_1 - \hat{P}_2)} = \frac{0.0091}{0.00175} = 5.2$
- $p - value = 0.0000001$

Step 4:

- Very strong evidence against H_0

t-test:

One-sample

Tests whether the mean of a normally distributed population is different from a specified value.
Null Hypothesis (H_0): states that the population mean is equal to some value (μ_0)
Alternate Hypothesis (H_A): states that the mean does not equal / is greater than / is less than μ_0
t-statistic: standardizes the difference between $\bar{x}$ and μ_0

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Degrees of freedom (df) = $n - 1$

Read the table of t-distribution critical values for p-value (probability that the sample mean was obtained by chance given μ_0 is the population mean) using the calculated t-statistic and degrees of freedom.

H_A: $\mu > \mu_0$ → the t-statistic is likely positive; read table as given

H_A : $\mu < \mu_0$ → the t-statistic is likely negative; the t-distribution is symmetrical so read the probability as if the t-statistic were positive

Note: if the t-statistic is of the 'wrong' sign, the p-value is 1 minus the p given in the chart

H_A: $\mu \neq \mu_0$ → read the p-value as if the t-statistic were positive and doubt it (to consider both less than and greater than)

If the p-value is less than the predetermined value for significance (called α and is usually 0.05), reject the null hypothesis and accept the alternate hypothesis.

Example:

You are experiencing hair loss and skin discoloration and think it might be because of selenium toxicity. You decide to measure the selenium levels in your tap water once a day for one week. Your results are given below. The EPA maximum containment level for safe drinking water is 0.05 mg/L. Does the selenium level in your tap water exceed the legal limit (assume α = 0.05)?

Day	Selenium mg/L
1	0.051
2	0.0505
3	0.049
4	0.0516
5	0.052
6	0.0508
7	0.0506

H_0: $\mu = 0.05$; H_A: $\mu > 0.05$;

Calculate the mean and standard deviation of your sample: $\bar{x} = 0.0508$

$$S^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{(0.051-0.0508)^2 + (0.0505-0.0508)^2 + etc\ ...}{6}$$
$$= 9.15 \times 10^{-7}$$

$$S = \sqrt{S^2} = 9.56 \times 10^{-4}$$

The t-statistic is: $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.0508-\ .05}{\frac{9.56\times10^{-4}}{\sqrt{7}}} = 2.17$ and the degrees of freedom are $n-1 = 7-1 = 6$

Looking at the t-distribution of critical values table, 2.17 with 6 degrees of freedom is between p = 0.05 and p = 0.025. This means that the p-value is less than 0.05, so you can reject H_0 and conclude that the selenium level in your tap water exceeds the legal limit.

t-test:

Two-sample

Tests whether the mean of two populations are significantly different form one another

Paired

- Each value of one group corresponds directly to a value in the other group; i.e., before and after values after drug treatment for each individual patient
- Subtract the two values for each individual to get one set of values (the differences) and use $\mu_0 = 0$ to perform a one-sample t-test

Unpaired

- The two populations are independent
- H_0: states that the means of the two populations are equal ($\mu_1 = \mu_2$)
- H_A: states that the means of the two populations are unequal or one is greater than the other ($\mu_1 \neq \mu_2,\ \mu_1 > \mu_2,\ \mu_1 < \mu_2$)

- t-statistic:

assuming equal variances: $t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$

assuming unequal variances: $t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$

degrees of freedom = $(n_1 - 1) + (n_2 - 1)$

- Read the table of t-distribution critical values for the p-value using the calculated t-statistic and degrees of freedom. Remember to keep the sign of the t-statistic clear (order of subtracting the sample means) and to double the p-value for an H_A of $\mu_1 \neq \mu_2$.

Example:

Consider the lifespan of 18 rats. 12 were fed a restricted calorie diet and lived an average of 700 days (standard deviation = 21 days). The other 6 had unrestricted access to food and lived an average of 668 days (standard deviation = 30 days). Does a restricted calorie diet increase the lifespan of rats (assume $\alpha = 0.05$)?

$\mu_1 = 700,\ s_1 = 21,\ n_1 = 12,\ \mu_2 = 668,\ s_2 = 30,\ n_2 = 6;$
$H_0: \mu_1 = \mu_2$;
$H_A: \mu_1 > \mu_2$ (because we are only asking if a restricted calorie diet increases
lifespan)
We cannot assume that the variances of the two populations are equal because
the different diets could also affect the variability on lifespan.
The t-statistic is: $t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} = \frac{700 - 668}{\sqrt{\frac{21^2}{12} + \frac{30^2}{6}}} = 2.342$
Degrees of freedom = $(n_1 - 1) + (n_2 - 1) = (12 - 1) + (6 - 1) = 16$
From the t-distributed table, the p-value falls between 0.01 and 0.02, so we do reject
H_0. The restricted calorie diet does increase the lifespan of rats.

z-test is used to statistically test a hypothesis when either we know the population variance, or we do not know the population variance but our sample size is large n ≥ 30. *If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.*

We perform a one-sample z-test when we want to compare a sample mean with the population mean and a two-sample z-test when we want to compare the means of two samples.

Example: A college claims that the students studying in their school are more intelligent than the average college. On calculating the IQ scores of 50 students, the average turns out to be 110. The mean of the population IQ is 100 and the standard deviation is 15. State whether the claim of principal is right or not at a 5% significance level.

First, we define the null hypothesis and the alternate hypothesis.
Our null hypothesis will be: $H_0: \mu = 100$
and our alternate hypothesis: $H_A: \mu > 100$

State the level of significance. Here, our level of significance given in this question (∝ =0.05), if not given then we take ∝=0.05.
Now, we look up to the z-table. For the value of ∝=0.05, the z-score for the right-tailed test is 1.645.
Now, we perform the Z-test on the problem:

$$Z = \frac{(X - \mu)}{(\sigma/\sqrt{n})}$$

Where:
X = 110
Mean (μ) = 100
Standard deviation (σ) = 15
Significance level (α) = 0.05
n = 50

$$Z = \frac{(110-100)}{(15/\sqrt{50})} = \frac{(10)}{2.12} = 4.71$$

Here 4.71 >1.645, so we reject the null hypothesis. If z-test statistics is less than z-score, then we will not reject the null hypothesis.

Chi-Square Test:

For Goodness of Fit

Checks whether or not an observed pattern of data fits some given distribution

H_0: the observed pattern fits the given distribution

H_a: the observed pattern does not fit the given distribution

The chi-square statistic is: $\chi^2 = \sum \frac{(O-E)^2}{E}$ (O is the observed value and E is the expected value)

Degree of freedom

$$= number\ of\ categories\ in\ the\ distribution - 1$$

Get the p-value from the table of χ^2 critical values using the calculated χ^2 and df values. If the p-value is less than α, the observed data does not fir the expected distribution. If $p > \alpha$, the data likely fits the expected distribution

Example 1:

You breed puffskeins and would like to determine the pattern of inheritance for coat color and purring ability. Puffskeins come in either pink or purple and either purr or hiss. You breed a purebred, pink purring male with a purebred, purple hissing female. All individuals of the F_1 generation are pink and purring. The F_2 offspring are shown below. Do the alleles for coat color and purring ability assort independently (assume $\alpha = 0.05$)?

Pink and Purring	Pink and Hissing	Purple and Purring	Purple and Hissing
143	60	55	18

Independent assortment means a phenotypic ratio of 9:3:3:1, so:

H_0: the observed distribution of F_2 offspring fits a 9:3:3:1 distribution

H_a: the observed distribution of F_2 offspring does not fit a 9:3:3:1 distribution

The expected values are:

Pink and Purring	Pink and Hissing	Purple and Purring	Purple and Hissing
155.25	51.75	51.75	17.25

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(143-155.25)^2}{155.25} + \frac{(60-51.75)^2}{51.75} + \frac{(55-51.75)^2}{51.75} + \frac{(81-17.25)^2}{17.25}$$

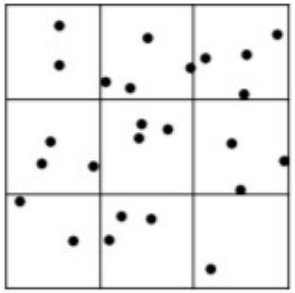
$$= 2.519$$

df = 4 – 1 = 3

From the table of χ^2 critical values, the p-value is greater than 0.25, so the alleles for coat color and purring ability do assort independently in puffskeins.

Example 2:

You are studying the pattern of dispersion of king penguins and the diagram below represents an area you sampled. Each dot is a penguin. Do the penguins display a uniform distribution (assume $\alpha = 0.05$)?



H_0: there is a uniform distribution of penguins

H_a: there is not a uniform distribution of penguins

There are a total of 25 penguins, so if there is a uniform distribution, there should be 2.778 penguins per square. There actual observed values are 2, 4, 4, 3, 3, 3, 2, 3, 1, so the χ^2 statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(1-2.778)^2}{2.778} + 2\left(\frac{(2-2.778)^2}{2.778}\right) + 4\left(\frac{(3-2.778)^2}{2.778}\right) + 2\left(\frac{(4-2.778)^2}{2.778}\right)$$

$$= 2.72$$

df = 9 – 1 = 8

From the table of χ^2 critical values, the p-value is greater than 0.25, so we do not reject H_0. The penguins do display a uniform distribution.

Chi-Square Test:

For Independence

Checks whether two categorical variables are related or not (independence)

H_0: the two variables are independent

H_a: the two variables are not independent

Does not make any assumptions about an expected distribution

The observed values $(\#_1, \#_2, \#_3, and\ \#_4)$ are usually presented as a table. Each row is a category of variable 1 and each column is a category of variable 2.

		Variable 1		Totals
		Category x	Category y	
Variable 2	Category a	$\#_1$	$\#_2$	$\#_1 + \#_2$
	Category b	$\#_3$	$\#_4$	$\#_3 + \#_4$
Totals		$\#_1 + \#_3$	$\#_2 + \#_4$	$\#_1 + \#_2 + \#_3 + \#_4$

The proportion of category x of variable 1 is the number of individuals in category x divided by the total number of individuals $\left(\frac{\#_1+\#_3}{\#_1+\#_2+\#_3+\#_4}\right)$. Assuming independence, the expected number of individuals that fall within category a of variable 2 is the proportion of category x multiplied by the number of individuals in category a $\left(\frac{\#_1+\#_3}{\#_1+\#_2+\#_3+\#_4}\right)(\#_1 + \#_2)$. Thus, the expected value is:

$$E = \frac{(\#_1 + \#_3)(\#_1 + \#_2)}{\#_1 + \#_2 + \#_3 + \#_4} = \frac{(row\ total)(column\ total)}{grand\ total}$$

Degree of freedom = $(r-1)(c-1)$ where r is the number of rows and c is the number of columns

The chi-square statistic is still $\chi^2 = \sum \frac{(O-E)^2}{E}$

Read the p-value from the table of χ^2 critical values.

Example: Given the data below, is there a relationship between fitness level and smoking habits (assume $\alpha = 0.05$)?

	Fitness Level				
	Low	Medium-Low	Medium-High	High	
Never smoked	113	113	110	159	495
Former smokers	119	135	172	190	616
1 to 9 cigarettes daily	77	91	86	65	319
≥ 10 cigarettes daily	181	152	124	73	530
	490	491	492	487	1960

H_0: fitness level and smoking habits are independent

H_a: fitness level and smoking habits are not independent

First, we calculate the expected counts. For the first cell, the expected count is:

$$E = \frac{(row\ total)(column\ total)}{grand\ total} = \frac{(495)(490)}{1960} = 123.75$$

	Fitness Level			
	Low	Medium-Low	Medium-High	High
Never smoked	123.75	124	124.26	122.99
Former smokers	154	154.31	154.63	153.06
1 to 9 cigarettes daily	79.75	79.91	80.08	79.26
≥ 10 cigarettes daily	132.5	132.77	133.04	131.69

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(113-123.75)^2}{123.75} + \frac{(113-124)^2}{124} + \frac{(110-124.26)^2}{124.26} + etc\ ...$$

$$= 91.73$$

df = $(r-1)(c-1) = (4-1)(4-1) = 9$

From the table of χ^2critical values, the p-value is less than 0.001, so we reject H_0 and conclude that there is a relationship between fitness level and smoking habits.

ANOVA (Analysis of Variance) is a statistical test used to analyse the difference between the means of more than two groups. A one-way ANOVA uses one independent variable while a two-way ANOVA uses two independent variables. The ANOVA table is set up as follows. Refer to a complete example here.

Source of Variation	Sums of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	F
Between Treatments	$SSB = \Sigma n_j \left(\bar{X}_j - \bar{X}\right)^2$	k-1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSE}$
Error (or Residual)	$SSE = \Sigma\Sigma \left(X - \bar{X}_j\right)^2$	N-k	$MSE = \frac{MSE}{N-k}$	
Total	$SST = \Sigma\Sigma \left(X - \bar{X}\right)^2$	N-1		

where

- X = individual observation,
- $\bar{X}_j$ = sample mean of the j^{th} treatment (or group),
- $\bar{X}$ = overall sample mean,
- k = the number of treatments or independent comparison groups, and
- N = total number of observations or total sample size.