Sample and sampling distribution facts summary I

- For several different statistics with sufficiently large sample size the sampling distribution is close to normal. We will be largely concerned with the the mean μ.
- This can be used to calculate estimates of several properties of population parameters.
- The sample mean $\bar{X}$ is an unbiased estimator $\hat{\mu}$ of the population mean μ. That is $\hat{\mu} = \bar{X}$.
- The sample variance s^2 (or S^2) is a biased estimate of the population variance σ^2. The unbiased estimator $\hat{\sigma}^2$ is $\frac{n}{n-1}s^2$. So, the sample variance s^2 slightly underestimates the population standard deviation.
- The square root of the variance estimator $\sqrt{\hat{\sigma}^2}$ written $\hat{\sigma}$ is actually a slightly biased estimator of σ (because of the non-linear operator $\sqrt{\ }$) but the bias is very small and in practice we pretend it is an unbiased estimator of σ.

Sample and sampling distribution facts summary II

- The variance of the sampling distribution (assuming it is normal) is ${\sigma^2}_{sd} = \frac{\sigma^2}{n}$, where n is sample size. The standard deviation of the sampling distribution, σ_{sd} is also called the standard error of the mean (or SEM). We will continue to use our notation, the subscript 'sd' on the symbol used for the population symbol σ. So $\sigma_{sd} = \frac{\sigma}{\sqrt{n}}$.

Confidence intervals, σ known

- Point estimates do not say how accurate they actually are.
- More useful to get a probability estimate for the interval within which we expect the population parameter will lie w.r.t to the random sample value of the statistic. The interval is called the **confidence interval** (CI) and the corresponding probability converted to a percentage is called the **confidence level** (CL). Popular values for CL are 90%, 95% and 99%. Note: *CL values be written as* α *and be in the interval* $[0, 1]$ *instead of* $[0, 100]$.
- The probability is associated with the interval w.r.t the random sample value. The actual population parameter is an unknown constant and has no associated probability. One way to think of CI-CL (in the frequency interpretation) is: If we draw a large number of samples then we expect the frequency with which the population value is within the CI w.r.t to the sample value to approach CL as the number of samples tends to infinity.

CI calculation

- The expression for the CI depends on the sampling distribution.
- Assuming the sampling distribution is normal we get the following CI:

$$\mu \in [\bar{X} - z^\star, \bar{X} + z^\star]$$

where $z^\star = \Phi^{-1}(1 - \frac{\alpha}{2}) = -\Phi^{-1}(\frac{\alpha}{2})$ and Φ is the cdf of the std. normal distribution.
The following table gives the $z^\star$ values (called critical values):

CL	90%	95%	99%
α	0.90	0.95	0.99
$z^\star$	1.645	1.96	2.576

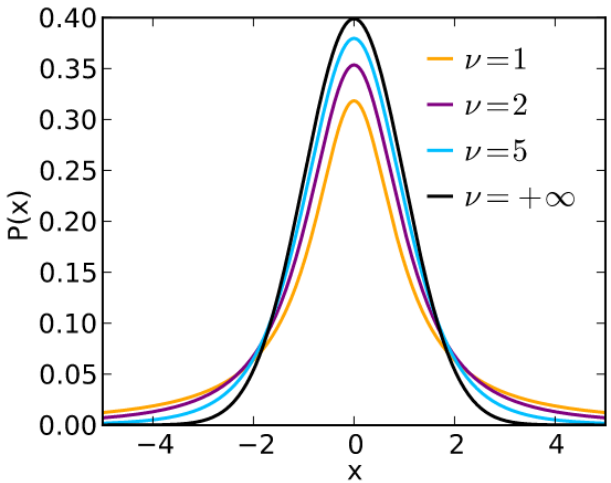
- Above translates to: $\bar{X} - z^\star \sigma_{sd} \leq \mu \leq \bar{X} + z^\star \sigma_{sd}$. Commonly used value $z^\star = 1.96$.

CI when σ not known

- σ is almost never known. Only s is known.
- When σ is not known we have to use sample std. deviation s and the t-distribution.

CI when σ not known

- σ is almost never known. Only s is known.
- When σ is not known we have to use sample std. deviation s and the t-distribution.



(source: wikipedia)

- The CI is: $\bar{X} - t^\star s_\mu \leq \mu \leq \bar{X} + t^\star s_\mu$ where $s_\mu = \frac{s}{\sqrt{n}}$, n: the sample size and $t^\star$ critical value for the CL chosen.

CL	90%	95%	99%
$t^\star$@20	1.73	2.093	2.861
$t^\star$@30	1.699	2.045	2.756
$z^\star$	1.645	1.96	2.576

CI for Difference of means, σ_1, σ_2 known

- Assume we have two populations and we measure the same attribute from random samples chosen independently from each population. Let the respective means be μ_1 and μ_2.
- if sample sizes are sufficiently large the sampling distribution of the difference of means is also close to normally distributed with mean $\bar{X}_d = \bar{X}_1 - \bar{X}_2$ (which is an unbiased estimator of $\mu_d = \mu_1 - \mu_2$) and $\sigma_{d_{sd}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
- So, the CI, as earlier is: $\bar{X}_d - z^\star \sigma_{d_{sd}} \leq \mu_d \leq \bar{X}_d + z^\star \sigma_{d_{sd}}$.

CI for Difference of means, σ_1, σ_2 unknown

- There are two cases: a) sample sizes are equal - n, b) unequal sample sizes respectively n_1, n_2.
- **Assumption**: the two variances are equal - homogeneity of variance.
- Equal sample sizes: $\bar{X}_d - t^\star \sigma_{d_{sd}} \leq \mu_d \leq \bar{X}_d + t^\star \sigma_{d_{sd}}$ where $\sigma_{d_{sd}} = \sqrt{\frac{2MSE}{n}}$, where $MSE = \frac{s_1^2 + s_2^2}{2}$, n: sample size. To find $t^\star$ the $DoF = 2n - 2$. Note: variance of sum of two independent RVs is a sum of the variances.
- Unequal samples: $\bar{X}_d - t^\star \sigma_{d_{sd}} \leq \mu_d \leq \bar{X}_d + t^\star \sigma_{d_{sd}}$ where $\sigma_{d_{sd}} = \sqrt{\frac{2MSE}{n_h}}$, $MSE = (SSE_1 + SSE_2)/\nu$, $\nu = n_1 + n_2 - 2$ is DoF, $n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}}$ and $SSE_i = \sum_{j=1}^{n_1} (x_j - \bar{X}_i)^2$, n_h is the harmonic mean and SSE_i are the sum of squared errors.

CI for proportion

- Notation: π population proportion, p sample proportion, n sample size.
- The CI: $p - z^\star \sigma_{p_{sd}} \leq \pi \leq p + z^\star \sigma_{p_{sd}}$ where $\sigma_{p_{sd}} = \sqrt{\frac{\pi(1-\pi)}{n}}$ when π is known (but this is being estimated and is not known), estimated by $\sqrt{\frac{p(1-p)}{n}}$.
- Continuity correction: A Gaussian is continuous, but proportions are not. Subtract $\frac{0.5}{n}$ from lower limit and add it to upper limit.
- **Assumptions**:
  - Sampling is random and independent.
  - Sufficient sample size - depends on p. Conservative rule of thumb: $np, n(1-p) \geq 10$.

CI for difference of proportions

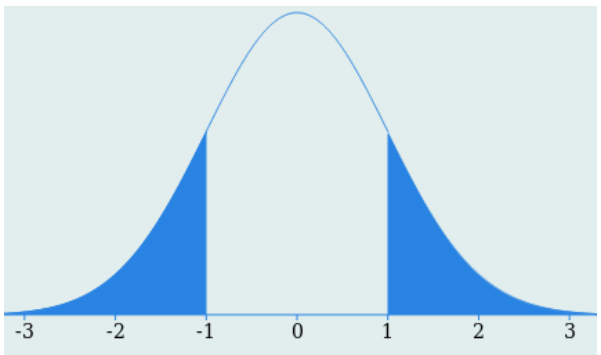
- Let $p_d = p_1 - p_2$. CI: $p_d - z^\star \sigma_{d_{sd}} \leq \pi_1 - \pi_2 \leq p_d + z^\star \sigma_{d_{sd}}$ when CLT applies - that is sufficiently large samples otherwise use $t^\star$ instead of $z^\star$. Here, $\sigma_{d_{sd}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
- The continuity correction: subtract $\frac{0.5 n_1 n_2}{n_1+n_2}$ from lower limit and add to upper limit.

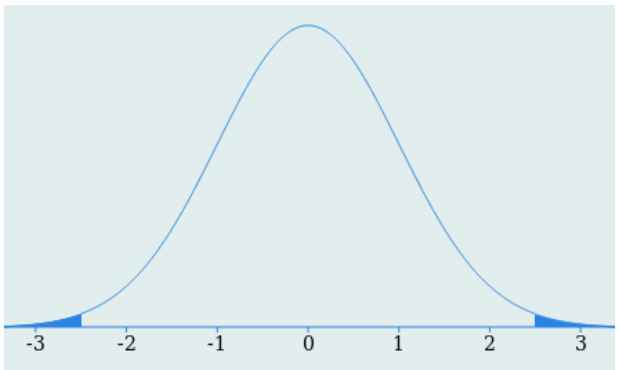
Hypothesis testing

- Hypothesis tests are techniques for making rational decisions in the context of incomplete information or uncertainty.
- It gives methods to decide whether claims about the behavioural effect(s) of one (or more) independent variable(s) can be believed.
- A hypothesis or claim is a precise statement about some population parameter(s) - like mean, median, proportion, difference of means, difference of proportions etc.
- Examples:
  - Average height of the IITK population is 155cm.
  - The amount of caffeine in a cup of coffee (150ml) will reduce the time students sleep in a class by at least 10 mins.
- Typical hypotheses: $\mu = \mu_0$, $\mu_1 = \mu_2$, $\mu_1 = \mu_2 = \mu_3$, $\pi = 0.4$, $\pi_1 - \pi_2 = 0$, $\rho_1 - \rho_2 = 0$ etc.

Logic of hypothesis testing

- How can we say that the independent variable(s) had an effect and it was not a matter of chance (or other factors)?
- If the probability that it is by chance is extremely small then one would tend to believe that the independent variable had an effect.





- What is extremely small is subjective. Typically, it is 5% or less or 1% or less. In behavioural studies the convention is 5%.

Null, alternate hypotheses

- The **null hypothesis** is the claim that there is no effect or difference. This is usually written H_0.
- The **alternate hypothesis** H_a is the negation of the null hypothesis and is the hypothesis the experimenter wishes to establish.
- H_a is viable or is accepted if H_0 can be shown to be not viable or rejected.
- The way this is done is to show that the sample statistic value makes H_0 extremely unlikely.
- Note that we avoid saying that H_a is true and H_0 is false. All that can be said is that H_a is highly probable and H_0 is extremely improbable.

P value, significance level, statistical significance

- Assuming H_0 is true the probability that the value of the statistic is as extreme or more extreme than the value actually observed is called the **P value** of the test. Note that P is not the probability of H_0 itself.
- The **significance level** (symbolized by α) is a pre-chosen value/level such that if $P \leq \alpha$ we will reject H_0.
- If $P \leq \alpha$ the results (or H_a) is **statistically significant** at level α. Note the close resemblance of the P-value to the CI.

Hypothesis testing works backwards

- To show that an effect exists (is highly probable) hypothesis testing works backwards:
  a) First it assumes that the effect does not exist - null hypothesis.
  b) Before data collection the rules for deciding whether the data is consistent with the null hypothesis are frozen - fix α.
  c) If the data are not consistent with the null hypothesis ($P \leq \alpha$) then our assumption that the effect does not exist (H_0) is rejected and therefore an effect (H_a) most probably exists.
- Not finding an effect is different from there is no effect.

One-sided, two-sided tests

- Most often tests are two-sided. For example $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$. For $\alpha = 0.5$ the critical/ rejection region are two tails covering probability mass of 0.025 respectively assuming a z-test and the critical value is 1.96.
- If the same is one sided where the test is one-sided i.e. $\mu_1 - \mu_2 < 0$ or $\mu_1 - \mu_2 > 0$ then the probability mass of 0.05 is only on one side and the critical value becomes 1.635 - so easier to reject the H_0.
- Many statisticians are wary of one sided tests since the effect in some sense is already known.

Statistical and practical significance

- A result may be statistically significant but practically unimportant or vice versa.

  Two diets are being tested for lowering cholesterol (mg/dl). $|\bar{X}_1 - \bar{X}_2| \geq 10$ have health effects. The table below shows some possible results.

No.	$\bar{X}_1 - \bar{X}_2$	$\sigma_{d_{sd}}$	t	P	$P < 0.05$?	95%CI	Pract. imp.
1	2	0.5	4	< 0.0001	Y	(1,3)	N
2	30	5	6	< 0.0001	Y	(20,40)	Y
3	30	14	2.1	0.032	Y	(2,58)	?
4	1	1	1	0.317	N	(-1,3)	N
5	2	30	0.1	0.947	N	(-58,62)	?
6	30	16	1.9	0.061	N	(-2,62)	?

Type I, Type II errors

- Two kinds of errors can arise a) A true null hypothesis is rejected and b) A false null hypothesis is not rejected. The table summarizes this:

	Actual state of H_0.	
Stat. inference	H_0 True	H_0 False
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

- α the significance level gives the probability of type I error.
- Hyp. testing is focused on reducing α. Possibly, because of widespread use in the bio-medical domain.
- The type II error is symbolized by β.