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 $E(\mu) = E(\hat{\mu}) = E(\bar{X})$. In practice we use $\bar{X}$.
- 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.

Sample and sampling distribution facts summary II

- 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.
- 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 $\sigma_{sd} = \frac{\sigma}{\sqrt{N}}$.

Point estimates I

- From a sample we usually want to estimate the population parameters.
- Two estimates can be calculated: point estimates and interval estimates.
- Let $P(X = x|\vec{\theta})$ be the pdf of some population, where $\vec{\theta}$ is the vector of population parameters. Example: (μ, σ^2) for normal or (n, p) for a binomial etc.
- Let $(X_1, \ldots, X_N)$ be an *iid* sample drawn from the population with values $(x_1, \ldots, x_N)$. Since the sample is *iid* the joint probability of seeing the sample give the population parameters is:
  $P(X_1 = x_1|\vec{\theta}) \times P(X_2 = x_2|\vec{\theta}) \times \ldots \times P(X_N = x_N|\vec{\theta})$
  This is called the **likelihood function** or just **likelihood** and written $\mathcal{L}(\vec{\theta}|\vec{x}) = \prod_{i=1}^{N} P(X = x_i|\theta)$.

Point estimates II

- One way to estimate $\vec{\theta}$ is to choose values of $\vec{\theta}$ that maximizes the likelihood. That is it maximizes the probability of jointly seeing the values actually seen - namely the likelihood.
- When the pdf is continuous this can be estimated by differentiating the likelihood w.r.t each parameter in $\vec{\theta}$.
- For the normal distribution this can be done analytically to obtain:
  $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{N}{N-1} s^2$ (corrected for bias).
- For likelihoods that are not differentiable numerical maximization has to be done using a search of the parameter space.

Properties of estimators

- An estimator $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta$.
- An estimator $\hat{\theta}$ is **consistent** if $\hat{\theta}$ is unbiased and $Var(\hat{\theta}) \to 0$ as $N \to \infty$.
- An estimator $\hat{\theta}_1$ is more **efficient** than another $\hat{\theta}_2$ if $\frac{\hat{\theta}_1}{\hat{\theta}_2} < 1$

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. This 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 are also written as* α *and can be in the interval* $[0, 1]$ *instead of* $[0, 100]$.
- The probability is associated with the interval when the value is from a random sample. The actual population parameter is an unknown constant and has no associated probability. Intuitively, think of CI-CL as the relative frequency (CL) with which the population value falls within the CI interval w.r.t the random sample value assuming we draw a large number of samples.

CI calculation I

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

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

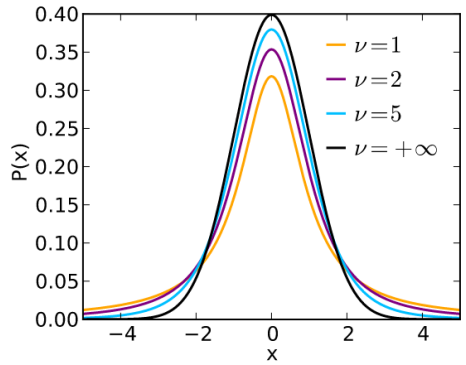
where $z^\star = \Phi^{-1}(1 - \frac{\alpha}{2}) = -\Phi^{-1}(\frac{\alpha}{2})$ and Φ is the cdf of the std. normal distribution.
The following table gives the $z^\star$ values (called critical values):

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

- For a normal distr. 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.



(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

The t distribution

- The Student-t or just t distribution family arises when estimating the mean of a normal population where sample size is small and population variance is not known.
- If sample size is N then we have $\nu = N - 1$ called degrees of freedom as an extra parameter in a t-distribution.
- The random variable $\frac{\bar{X}-\mu}{s/\sqrt{N}}$ has a t-distribution with $\nu = N - 1$ where $X \sim N(\mu, \sigma^2)$, s is the sqrt. of the unbiased sample variance and N is the sample size.

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. When 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

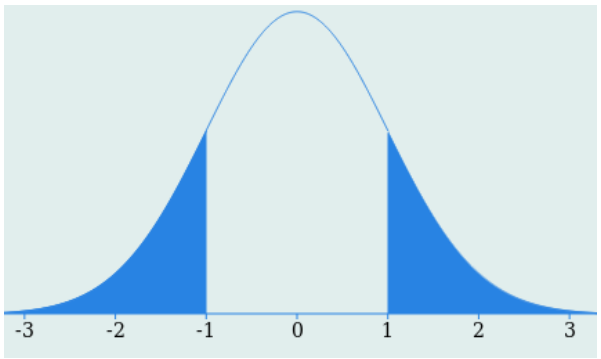
- Let $p_d = p_1 - p_2$. CI: $p_d - z^\star \sigma_{d_{sd}} \leq \pi_1 - \pi_2 \leq p_d + z^\star \sigma_{d_{sd}}$ when CLT applies - that is sufficiently large samples otherwise use $t^\star$ instead of $z^\star$. Here, $\sigma_{d_{sd}} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$.
- The 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%.