MATH2361: Probability and Statistics 2024-25

Instructor: **Dr. Harsha K V**, Department of Mathematics

GITAM Hyderabad

Formula list for Unit 4

I. Important Formulas

- Consider a **Random sample** $X_1, \cdots, X_n$ from a population with mean μ, and variance σ^2. Then average $\bar{X}$ is also a random variable, and we have

$$E(\bar{X}) = \mu.$$

For infinite population, variance of $\bar{X}$ is

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

For finite population with size N, variance of $\bar{X}$ is

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

- **Finite Population Correction Factor (FPC)=** $\frac{N-n}{N-1}$.
- **Standard error (S.E.)** of an estimator = standard deviation of the estimator
- Standard error of average $\bar{X}$, and sample proportion $\hat{p}$ are

$$\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- **Maximum error** E of sample proportion:

$$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- Maximum sample size n for at most E maximum error,

$$n \approx \frac{1}{4}\left(\frac{Z_{\alpha/2}}{E}\right)^2$$

- **Central Limit Theorem (CLT)**: Consider a random sample $X_1, \cdots, X_n$ from a population having mean $E(X) = \mu$, and finite variance $\text{Var}(X) = \sigma^2$. Then the random variable
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
converges in distribution to the standard normal distribution $N(0, 1)$ as n tends to ∞.

Some Useful Distributions

1) **Chi-square distribution with degrees of freedom (df),** χ^2_{df}: Let $Z_1, \cdots, Z_k$ are independent $N(0, 1)$ random variables. Then random variable $Y = Z_1^2 + \cdots + Z_k^2$ has the chi-square distribution with parameter $df = k$, denoted by χ^2_k. Its not symmetric around origin.
2) **Student's t-distribution with with degrees of freedom (df),** t_{df}: Let Z be a standard normal random variable $N(0, 1)$, Y be a chi-square random variable χ^2_k and Z, and Y are independent, then the random variable $t = \frac{Z}{\sqrt{Y/k}}$ follows t-distribution t_k with $df = k$.
3) **F-distribution with parameters** k_1, k_2**,** F_{k_1,k_2}: Let Y_1, Y_2 be two independent chi-square random variables with df's, k_1 and k_2. Then random variable $F = \frac{Y_1/k_1}{Y_2/k_2}$ has F-distribution with parameters k_1, k_2, denoted by F_{k_1,k_2}.

Sampling Distributions

- If population is normal, and σ is known, the $\frac{(n-1)S^2}{\sigma^2}$ follows chi-square distribution χ^2_{n-1}.
- If S_1^2 and S_2^2 are variances of two independent samples with sizes n_1, n_2 from two normal populations, then $\frac{S_1^2}{S_2^2}$ has F-distribution with parameters, $k_1 = n_1 - 1, k_2 = n_2 - 1$.

Hypothesis Testing

- **Null hypothesis** H_0: It is the status-quo claim about the population parameter. Null hypothesis is often the default or established belief about the population parameter such as there is no effect or any difference between the variables being considered.
- **Alternative hypothesis** H_1: It is a statement that proposes that there is an effect, a difference, or a relationship between the variables being studied. It is the hypothesis that you want to test against the null hypothesis.
- **Type-I error**: Occurs when rejecting a true null hypothesis
- **Type-II error**: Occurs fail to reject a false null hypothesis
- **Significance level** is the probability of Type-I error, denoted by α.
- The probability of a Type II error is denoted by β. Power of test is defined as $1 - \beta$.
- The **rejection region** (also called the **critical region**) is the range of values of the test statistic which leads to rejection of the null hypothesis.

- **Acceptance region** is the range of values of the test statistic for which we fail to reject the null hypothesis.
- **Critical value**: A point of the test statistic distribution that helps you decide whether the observed test statistic falls in the rejection region of H_0.

Z-test Critical Values and Rejection criteria for H_0:

H_1	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Two-tailed	$Z > 2.58$ or $Z < -2.58$	$Z > 1.96$ or $Z < -1.96$	$Z > 1.645$ or $Z < -1.645$
Right-tailed	$Z > 2.33$	$Z > 1.645$	$Z > 1.28$
Left-tailed	$Z < -2.33$	$Z < -1.645$	$Z < -1.28$

List of formulas for $(1-\alpha)100\%$ Confidence interval (C.I.)

	$(1-\alpha)100\%$ **Confidence interval (C.I.) or Interval estimates**
1	For $n \geq 30$: If σ known, **C.I for mean** μ is $\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$ Here, $\bar{x}$ is the value of $\bar{X}$. If σ unknown, then use sample standard deviation value s, and C.I. is $\bar{x} - Z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{s}{\sqrt{n}}.$
2	For $n < 30$: For normal population with σ known: **C.I for mean** μ is $\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$ For normal population with σ unknown: **C.I for mean** μ is $\bar{x} - t_{df,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{df,\alpha/2}\frac{s}{\sqrt{n}}.$ $t_{df,\alpha/2}$ is the critical value from t-distribution with $df = n-1$.
3	For $n_1, n_2 \geq 30$: If σ_1, σ_2 are known: **C.I for** $\delta = \mu_1 - \mu_2$ is $(\bar{x} - \bar{y}) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x} - \bar{y}) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$ If σ_1, σ_2 are unknown, **C.I for** δ is $(\bar{x} - \bar{y}) - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x} - \bar{y}) + Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$
4	For $n_1, n_2 < 30$, and both populations are normal: If $\sigma_1 = \sigma_2 = \sigma$, and is unknown: **C.I for** $\delta = \mu_1 - \mu_2$ is $(\bar{x} - \bar{y}) - t_{df,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \mu_1 - \mu_2 < (\bar{x} - \bar{y}) + t_{df,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where s_p^2 is the pooled variance given by $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}, \quad df = n_1 + n_2 - 2.$

List of formulas for $(1-\alpha)100\%$ Confidence interval (C.I.)

	$(1-\alpha)100\%$ **Confidence interval (C.I.)**
5	For $n_1, n_2 < 30$, and both populations are normal: If $\sigma_1 \neq \sigma_2$, and both are unknown: **C.I for** $\delta = \mu_1 - \mu_2$ is $(\bar{x}-\bar{y}) - t_{df,\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)} < \mu_1 - \mu_2 < (\bar{x}-\bar{y}) + t_{df,\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}$ where df is estimated as (take only integer part) $df \approx \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{(\frac{s_1^2}{n_1})^2}{n_1-1}+\frac{(\frac{s_2^2}{n_2})^2}{n_2-1}\right)}.$
6	For $n < 30$: For matched pair (X, Y) with $D = X - Y$ is normal, and σ_D unknown: **C.I for mean** μ_D is $\bar{D} - t_{df,\alpha/2}\frac{s_D}{\sqrt{n}} < \mu < \bar{D} + t_{df,\alpha/2}\frac{s_D}{\sqrt{n}}.$ where $df = n - 1$, and s_D is the standard deviation of D.
7	For $n \geq 2$, and population is normal: **C.I for** σ^2 is $\frac{(n-1)s^2}{\chi^2_{df,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{df,\alpha/2}}.$ **C.I for** σ is $\sqrt{\frac{(n-1)s^2}{\chi^2_{df,1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{df,\alpha/2}}}$ where $\chi^2_{df,\alpha/2}$, and $\chi^2_{df,1-\alpha/2}$ are critical values from χ^2-distribution, where $df = n - 1$.
8	For $n \geq 30$: **C.I for population proportion** p is $\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is the sample proportion $\hat{p} = \frac{X}{n}$, where X is the number of “successes” in n samples.

Large Sample Hypothesis Tests (Z-tests)

Large sample means $n \geq 30$, and $n_1, n_2 \geq 30$: Level of significance α

Assumptions: Any population with finite variance

1

Test for Single mean: $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ (or $\mu > \mu_0$ or $\mu < \mu_0$)

1) If σ known, test statistic under H_0 is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

2) If σ unknown, then use sample standard deviation S, test statistic under H_0 is

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Test for Single Population Proportion: $H_0 : p = p_0$ against $H_1 : p \neq p_0$ (or $p > p_0$ or $p < p_0$)

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \text{sample proportion} \quad \hat{p} = \frac{X}{n}$$

where X is the number of "successes" in n samples.

In all cases, **Decision rule**: reject H_0 if

- $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$, for two-tailed $H_1 : \mu \neq \mu_0$ (For proportion $H_1 : p \neq p_0$)
- $Z > Z_\alpha$ for right-tailed $H_1 : \mu > \mu_0$ (For proportion $H_1 : p > p_0$)
- $Z < -Z_\alpha$ for left-tailed $H_1 : \mu < \mu_0$ (For proportion $H_1 : p < p_0$).

2

Test for difference of means of two populations: $H_0 : \mu_1 - \mu_2 = \delta_0$ against $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (or $\mu_1 - \mu_2 > \delta_0$ or $\mu_1 - \mu_2 < \delta_0$)

1) If σ_1, σ_2 are known: test statistic under H_0 is

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

2) If σ_1, σ_2 are unknown, test statistic under H_0 is

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where S_1, S_2 are sample standard deviations for two populations.

Decision rule: reject H_0 if

- $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$, for two-tailed $H_1 : \mu_1 - \mu_2 \neq \delta_0$
- $Z > Z_\alpha$ for right-tailed $\mu_1 - \mu_2 > \delta_0$
- $Z < -Z_\alpha$ for left-tailed $\mu_1 - \mu_2 < \delta_0$.

Small Sample Hypothesis Tests (t-tests)

	Small sample $n < 30$, and $n_1, n_2 < 30$: Level of significance α
	Assumptions: All populations are normal, and population variances are unknown
1	**Test for Single mean:** $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ (or $\mu > \mu_0$ or $\mu < \mu_0$) Here, σ is unknown: Test statistic under H_0 is $t = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}, \quad \text{where} \quad S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ **Matched pair t-test:** For related (X, Y), let $D = X - Y$, and D is normal, with mean μ_D, and population variance σ_D^2 unknown: $H_0 : \mu_D = 0$ against $H_1 : \mu_D \neq 0$ (or $\mu_D > 0$ or $\mu_D < 0$) Test statistic under H_0 is $t = \dfrac{\bar{D}}{S_D/\sqrt{n}},$ where $\bar{D}$ is the sample mean, and S_D is standard deviation of n observations of D. In both cases, **Decision rule**: reject H_0 if • $t > t_{df,\alpha/2}$ or $t < -t_{df,\alpha/2}$, for two-tailed $H_1 : \mu \neq \mu_0$ (For matched t-test $H_1 : \mu_D \neq 0$) • $t > t_{df,\alpha}$ for right-tailed $H_1 : \mu > \mu_0$ (For matched t-test $H_1 : \mu_D > 0$) • $t < -t_{df,\alpha}$ for left-tailed $H_1 : \mu < \mu_0$ (For matched t-test $H_1 : \mu_D < 0$) where $df = n - 1$.
2	**Test for difference of means of two independent populations:** $H_0 : \mu_1 - \mu_2 = \delta_0$ against $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (or $\mu_1 - \mu_2 > \delta_0$ or $\mu_1 - \mu_2 < \delta_0$) 1) If we have prior information that $\sigma_1 = \sigma_2 = \sigma$ (but σ unknown): test statistic under H_0 is $t = \dfrac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}, \quad \text{where} \quad S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$ Here, t has sampling distribution t_{df}, where $df = n_1 + n_2 - 2$. 2) If we have prior information that $\sigma_1 \neq \sigma_2$, test statistic under H_0 is $t = \dfrac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \rightarrow \quad \text{sampling distribution } t_{df}, \quad df \approx \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1} \right)}.$ In both cases **Decision rule**: reject H_0 if • $t > t_{df,\alpha/2}$ or $t < -t_{df,\alpha/2}$, for two-tailed $H_1 : \mu_1 - \mu_2 \neq \delta_0$ • $t > t_{df,\alpha}$ for right-tailed $\mu_1 - \mu_2 > \delta_0$ • $t < -t_{df,\alpha}$ for left-tailed $\mu_1 - \mu_2 < \delta_0$. Note that df is different in both cases, so choose as mentioned above!

Other Hypothesis Tests

	Sample size $n \geq 2$ and Level of significance α
	Assumptions: All populations are normal, and observations are independent
1	**Test for Single mean:** $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ (or $\mu > \mu_0$ or $\mu < \mu_0$) (σ is known): Test statistic under H_0 is $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$ Here, Z has the sampling distribution $N(0, 1)$. **Decision rule**: reject H_0 if • $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$, for two-tailed $H_1 : \mu \neq \mu_0$ • $Z > Z_\alpha$ for right-tailed $H_1 : \mu > \mu_0$ • $Z < -Z_\alpha$ for left-tailed $H_1 : \mu < \mu_0$.
2	**Test for variance of population** $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$ (or $\sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$) Test statistic under H_0 is $Y = \frac{(n-1)S^2}{\sigma_0^2}$ where S^2 is the sample variance. Here, Y has the sampling distribution chi-square χ^2_{df}, where $df = n - 1$. **Decision rule**: reject H_0 if • $Y > \chi^2_{df,\alpha/2}$ or $Y < \chi^2_{df,1-\alpha/2}$, for two-tailed $H_1 : \sigma^2 \neq \sigma_0^2$ • $Y > \chi^2_{df,\alpha}$ for right-tailed $\sigma^2 > \sigma_0^2$ • $Y < \chi^2_{df,1-\alpha}$ for left-tailed $\sigma^2 < \sigma_0^2$.