Math 3020, Probability and Statistics for CSC

Final Exam Formulas

- Probability of the Sample Space:
$$P(\Omega) = 1$$
- Probability of the Empty Set:
$$P(\varnothing) = 0$$
- Complement Rule for any event A:
$$P(A^c) = 1 - P(A)$$
- Addition Rule for Two Arbitrary Events:
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
- Addition Rule for Three Arbitrary Events:
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$
- Multiplication Rule for Two Independent Events:
$$P(A \cap B) = P(A)P(B)$$
- Formula for "Permutations with Replacement":
$$P_r(n,k) = n \cdot n \cdot n ... \cdot n = n^k$$
- Formula for "Permutations without Replacement":
$$P(n,k) = \frac{n!}{(n-k)!} = n(n-1)(n-2)(n-3)...(n-k+1)$$
- Formula for "Combinations without Replacement":
$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$
- Formula for "Combinations with Replacement":
$$C_r(n,k) = \binom{k+n-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$
- Multiplication Rule for general events:
$$P(A \cap B) = P(A|B) \cdot P(B)$$
- Bayes Rule:
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
- The Expectation is defined as:
$$\mu = \mathbf{E}(X) = \sum_x xP(x)$$
- The Variance is defined as:
$$\sigma^2 = \mathrm{Var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \sum_x (x-\mu)^2 P(x),$$

- **Bernoulli Distribution** with the probability of success p:
  - The pmf is
  $$P(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$
  - The expectation is
  $$\mu = \mathbf{E}(X) = p.$$
  - The variance is:
  $$\sigma^2 = \mathrm{Var}(X) = p(1-p) = pq.$$

- **Binomial Distribution** with n trials and probability of success p:
  - The pmf is:
  $$P(x) = \mathbf{P}\{X = x\} = \binom{n}{x} p^x q^{n-x}, \quad \text{for } x = 0, 1, 2, ..., n.$$
  - The expectation is:
  $$\mu = \mathbf{E}(X) = np.$$
  - The variance is:
  $$\sigma^2 = \mathrm{Var}(X) = npq.$$

- **Geometric Distribution** with the probability of success p:
  - The pmf is:
  $$P(x) = \mathbf{P}\{X = x\} = (1-p)^{x-1} p, \quad \text{for } x = 1, 2, ...$$
  - The expectation is:
  $$\mu = \mathbf{E}(X) = \frac{1}{p}$$
  - The variance is:
  $$\sigma^2 = \mathrm{Var}(X) = \frac{1-p}{p^2}$$

- **Poisson Distribution** with the frequency λ:
  - The pmf is:
  $$P(x) = \mathbf{P}\{X = x\} = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, \quad \text{for} \quad x = 0, 1, 2, ...$$
  - The expectation is:
  $$\mu = \mathbf{E}(X) = \lambda$$
  - The variance is:
  $$\sigma^2 = \mathrm{Var}(X) = \lambda$$

- To solve the probability of Binomial Distributions using R:
  - $\mathbf{P}\{X = a\} = \mathrm{dbinom}(a, n, p)$
  - $\mathbf{P}\{X \leq a\} = \mathrm{pbinom}(a, n, p)$
  - $\mathbf{P}\{X > a\} = 1 - \mathrm{pbinom}(a, n, p)$
  - $\mathbf{P}\{X \geq a\} = 1 - \mathrm{pbinom}(a-1, n, p)$

- To solve the probability of Poisson Distributions using R:
  - $\mathbf{P}\{X = a\} = \text{dpois}(a, \lambda)$
  - $\mathbf{P}\{X \leq a\} = \text{ppois}(a, \lambda)$
  - $\mathbf{P}\{X > a\} = 1 - \text{ppois}(a, \lambda)$
  - $\mathbf{P}\{X \geq a\} = 1 - \text{ppois}(a - 1, \lambda)$

- **Uniform Distribution** on an interval (a, b):
  - The pdf is
  $$f(x) = \frac{1}{b-a} \quad \text{for} \quad a < x < b$$
  (and $f(x) = 0$ otherwise).
  - The expectation is:
  $$\mu = \mathbf{E}(X) = \frac{a+b}{2}$$
  - The variance is:
  $$\sigma^2 = \text{Var}(X) = \frac{(b-a)^2}{12}$$

- **Exponential Distribution** with frequency λ:
  - The pdf is
  $$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad x > 0$$
  (and $f(x) = 0$ otherwise).
  - The cdf is
  $$F(x) = 1 - e^{-\lambda x} \quad \text{for} \quad x > 0$$
  (and $F(x) = 0$ otherwise).
  - The expectation is:
  $$\mu = \mathbf{E}(X) = \frac{1}{\lambda}$$
  - The variance is:
  $$\sigma^2 = \text{Var}(X) = \frac{1}{\lambda^2}$$

- **Normal Distribution** with the mean μ and standard deviation σ:
  - The pdf is
  $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$
  - The cdf is
  $$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt$$
  - The expectation is:
  $$\mathbf{E}(X) = \mu$$
  - The variance is:
  $$\text{Var}(X) = \sigma^2$$

- To solve the probability of Standard Normal Distribution (i.e., $Z \sim N(0,1)$) using R :

  * $\mathbf{P}\{Z \leq z\} = \mathbf{P}\{Z < z\} = \text{pnorm}(z)$
  * $\mathbf{P}\{Z \geq z\} = \mathbf{P}\{Z > z\} = 1 - \text{pnorm}(z)$
  * $\mathbf{P}\{z_1 \leq Z \leq z_2\} = \text{pnorm}(z_2) - \text{pnorm}(z_1)$

- **Central Limit Theorem (informally)**

  Let $X_1, X_2, X_3, \ldots$ be independent random variables with the same expectation $\mu = \mathbf{E}(X_i)$ for all i and with the same standard deviation $\sigma = \text{Std}(X_i)$ for all i, and let

  $$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \ldots + X_n$$

  As $n \to \infty$, the sum S_n follows a Normal Distribution:

  $$S_n \sim \text{Normal}(n\mu, \sigma\sqrt{n})$$

- Summary of Discrete distributions and Continuous distributions:

Distribution:	**Discrete**	**Continuous**
Defined by:	$P(x) = \mathbf{P}\{X = x\}$ (pmf)	$f(x) = F'(x)$ (pdf)
Calculating probabilities:	$\mathbf{P}\{a < X \leqslant b\} = \sum_{a < x_j \leqslant b} P(x_j)$	$\mathbf{P}(a < X \leqslant b\} = \int_a^b f(x)\,\mathrm{d}x$
Cumulative distribution function (cdf):	$F(x) = \sum_{x_j \leqslant x} P(x_j)$	$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$
Total probability:	$\sum_{x_j} P(x_j) = 1$	$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$
Expectation:	$\mu = \mathbf{E}(X) = \sum_{x_j} x_j P(x_j)$	$\mu = \mathbf{E}(X) = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x$
Variance:	$\sigma^2 = \text{Var}(X) = \sum_{x_j} (x_j - \mu)^2 P(x_j)$	$\sigma^2 = \text{Var}(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$

- z-score in R:

  $$z_{\alpha/2} = \text{qnorm}(1 - \alpha/2)$$
  $$z_{\alpha} = \text{qnorm}(1 - \alpha)$$

- t-score in R:

  $$t_{\alpha/2} = \text{qt}(1 - \alpha/2, \text{df})$$
  $$t_{\alpha} = \text{qt}(1 - \alpha, \text{df})$$

- The confidence interval for **population mean** with known σ:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \left[\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\quad \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

- The confidence interval for the difference between **two population means**:

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

- Sample size needed for a given precision for **population mean**:

In order to attain a margin of error Δ for estimating a **population mean** with a confidence level $(1-\alpha)$, a sample size is required:

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{\Delta}\right)^2$$

- Sample size needed for a given precision for **population proportion**:

In order to attain a margin of error Δ for estimating a **population proportion** with a confidence level $(1-\alpha)$, a sample size is required:

$$n \geq 0.25 \left(\frac{z_{\alpha/2}}{\Delta}\right)^2$$

Alternatively, we can generate to any proportion if the proportion p is given:

$$n \geq p(1-p) \left(\frac{z_{\alpha/2}}{\Delta}\right)^2$$

- The confidence interval for **population proportion**:

$$\widehat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

- The confidence interval for the difference of **population proportions**:

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

- The confidence interval for the **population Mean** using t-value:

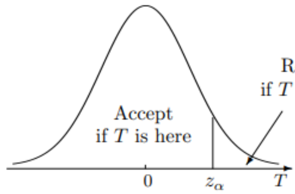
$$\overline{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation and $t_{\alpha/2}$ is a critical value from t-distribution with $n-1$ degrees of freedom.

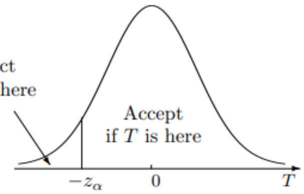
- Possible Errors:

	Result of the test	
	Reject H_0	**Accept H_0**
H_0 is true	*Type I error*	correct
H_0 is false	correct	*Type II error*

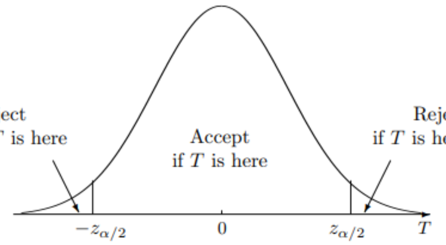
- Rejection Regions:



(a) Right-tail Z-test



(b) Left-tail Z-test



(c) Two-sided Z-test

- Z-tests are summarized in the table below.

Null hypothesis	Parameter, estimator	If H_0 is true:		Test statistic
H_0	θ, $\widehat{\theta}$	$\mathbf{E}(\widehat{\theta})$	$\mathrm{Var}(\widehat{\theta})$	$Z = \dfrac{\widehat{\theta} - \theta_0}{\sqrt{\mathrm{Var}(\widehat{\theta})}}$
One-sample Z-tests for means and proportions, based on a sample of size n				
$\mu = \mu_0$	μ, $\overline{X}$	μ_0	$\dfrac{\sigma^2}{n}$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$
$p = p_0$	p, $\widehat{p}$	p_0	$\dfrac{p_0(1-p_0)}{n}$	$\dfrac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y = D$	$\mu_X - \mu_Y$, $\overline{X} - \overline{Y}$	D	$\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$	$\dfrac{\overline{X} - \overline{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
$p_1 - p_2 = D$	$p_1 - p_2$, $\widehat{p}_1 - \widehat{p}_2$	D	$\dfrac{p_1(1-p_1)}{n} + \dfrac{p_2(1-p_2)}{m}$	$\dfrac{\widehat{p}_1 - \widehat{p}_2 - D}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}}$
$p_1 = p_2$	$p_1 - p_2$, $\widehat{p}_1 - \widehat{p}_2$	0	$p(1-p)\left(\dfrac{1}{n} + \dfrac{1}{m}\right)$, where $p = p_1 = p_2$	$\dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$ where $\widehat{p} = \dfrac{n\widehat{p}_1 + m\widehat{p}_2}{n+m}$

- T-tests are summarized in the table below.

Hypothesis H_0	Conditions	Test statistic t	Degrees of freedom
$\mu = \mu_0$	Sample size n; unknown σ	$t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown but equal standard deviations, $\sigma_X = \sigma_Y$	$t = \dfrac{\overline{X} - \overline{Y} - D}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$	$n + m - 2$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown, unequal standard deviations, $\sigma_X \neq \sigma_Y$	$t = \dfrac{\overline{X} - \overline{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Satterthwaite approximation, formula (9.12)

- R code for 95% Confidence Intervals for one Proportion:

  prop.test(c(x,0), c(n,n), conf.level= 0.95, correct = F)

- R code for 95% Confidence Intervals for two Proportions:

  prop.test(c(x1,x2), c(n1,n2), conf.level= 0.95, correct = F)

- R code to find the critical $t_{\alpha/2}$:

$$t_{\alpha/2} = \mathbf{qt}(1 - \tfrac{\alpha}{2},\ df)$$

- R code for 95% Confidence Intervals when sample standard deviation (s) is given:

  t.test(x, conf.level = 0.95)

  where, x=c(x1, x2, x3, ..., xn).

- R code to perform a hypothesis test for one proportion:

  prop.test(x, n, p, alternative = "greater" (or "less", or "two.sided"), correct = F)

- R code to perform a hypothesis test for two proportions:

  prop.test(c(x1,x2), c(n1,n2), alternative = "greater" (or "less", or "two.sided"), correct = F)

- R code to perform a hypothesis test for one mean if we have the full data set:

  t.test(x, mu = ??, alternative = "greater" (or "less", or "two.sided"))

- R code to find the probability using t-test:

$$\mathbf{P}\{t < a\} = \text{pt(a, df)}$$
$$\mathbf{P}\{t > b\} = \text{1 - pt(b, df)}$$