Probability Theory

Consider events A and B.

- Rule 1: $0 \leq Pr\{A\} \leq 1$.
- Rule 2: If there are k events $A_1, A_2, \ldots, A_k$ that make up all possible events, then $\sum_{i=1}^{k} Pr\{A_i\} = 1$
- Rule 3: The probability that A does not occur is: $P\{A^C\} = 1 - Pr\{A\}$
- Rule 4: For any two events A and B, the probability of "A occurs or B occurs or both occur" is: $P\{A \text{ or } B\} = Pr\{A\} + Pr\{B\} - Pr\{A \text{ and } B\}$
- Rule 5: If A and B are mutually exclusive (or disjoint), then $Pr\{A \text{ and } B\} = 0$
- Rule 6: The conditional probability of A given B has occurred is: $Pr\{A|B\} = \frac{Pr\{A \text{ and } B\}}{Pr\{B\}}$
- Rule 7: $Pr\{A \text{ and } B\} = Pr\{A|B\}Pr\{B\}$
- Rule 8: $Pr\{A \text{ and } B^c\} = Pr\{A\} - Pr\{A \text{ and } B\}$
- Rule 9: $Pr\{A^C|B\} = 1 - Pr\{A|B\}$
- Rule 10: If an event A is split by multiple events $B_1, B_2, \ldots, B_k$, then the following is true: $Pr\{A\} = Pr\{A \text{ and } B_1\} + Pr\{A \text{ and } B_2\} + \cdots + Pr\{A \text{ and } B_k\}$

  For two events A and B:
  $Pr\{A\} = Pr\{A \text{ and } B\} + Pr\{A \text{ and } B^C\}$
- For two events A and B which are independent, both of the following properties hold true:
  1. $Pr\{A \text{ and } B\} = Pr\{A\}Pr\{B\}$
  2. $Pr\{A|B\} = Pr\{A\}$

Binomial Random Variables

If Y is a binomial random variable;

- $Pr\{Y = j\} = \binom{n}{j} p^j (1-p)^{n-j}$
  where $\binom{n}{j} = \frac{n!}{j!(n-j)!}$
- $\mu_Y = np$
- $\sigma_Y^2 = np(1-p)$

Normal Random Variables

If Y is a normal random variable with mean μ_Y, standard deviation σ_Y (i.e $Y \sim N(\mu_Y, \sigma_Y)$) then;

- $Z = \frac{Y - \mu_Y}{\sigma_Y}$ is standard normal, i.e. $Z \sim N(0,1)$.
- $Pr\{Z > a\} = 1 - P(Z < a)$
  for some constant a.
- $Pr\{a < Z < b\} = Pr\{Z < b\} - Pr\{Z < a\}$
  for some constants a and b.
- The k^{th} percentile of Y is :
  $Y^{(k)} = \mu_Y + Z^{(k)}\sigma_Y$
  where $Z^{(k)}$ is the k^{th} percentile of a Z.

Distribution of the Sample Mean

If a random sample from a population Y with mean μ_Y and standard deviation σ_Y is taken and either

(i): The population is normally distributed, or

(ii): $n \geq 30$

then the sample mean $\bar{Y}$ is normally distributed with mean μ_Y, standard deviation $\frac{\sigma_Y}{\sqrt{n}}$ (I.e., $\bar{Y} \sim N(\mu_Y, \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}})$).

Confidence Interval for μ

- A (1-α)100% CI for μ is:
  $\bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$
  at d.f. $= n - 1$
- To calculate what sample size you should take for a margin of error within e:
  $n = \frac{t_{\alpha/2}^2 s^2}{e^2}$
  where we use d.f. $= \infty$ for $t_{\alpha/2}$.

Confidence Interval for $\mu_1 - \mu_2$

- A (1- α)100% CI for $\mu_1 - \mu_2$ is:
  $(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  at d.f. $= \nu$ (this will be given).

Hypothesis Test for $\mu_1 - \mu_2$

- Step 1: State the null and alternative.
- Step 2: The test-statistic is: $t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
  at d.f. $= \nu$ (this will be given)
- Step 3: The possible p-values are:
  If $H_A : \mu_1 \neq \mu_2$, p-value $= 2Pr\{t > |t_s|\}$
  If $H_A : \mu_1 < \mu_2$, p-value $= Pr\{t < t_s\}$
  If $H_A : \mu_1 > \mu_2$, p-value $= Pr\{t > t_s\}$

General Definitions

- p-value: The probability of observing our sample data or more extreme, if the null hypothesis is true.
- Type I error: When we reject the null, if in reality the null is true.
- Type II error: When we fail to reject the null, if in reality the null is false.
- Step 4: Decision Rule (for any hypothesis test):
  If p-value $< \alpha$, reject H_0.
  If p-value $\geq \alpha$, fail to reject H_0.

Definitions for ANOVA

- n_i = sample size for group i
  $n_\bullet$ = overall sample size = $\sum_{i=1}^{I} n_i$

- $\bar{y}_i$ = sample mean for group i
  $\bar{\bar{y}} = \frac{\sum_{i=1}^{I} n_i \bar{y}_i}{n_\bullet}$ = overall group mean
  s_i = sample standard deviation for group i
  I = total number of groups

- $SSB = \sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$
  $d.f.\{B\} = I - 1$
  $MSB = \frac{SSB}{d.f.\{B\}}$

- $SSW = \sum_{i=1}^{I} (n_i - 1)s_i^2$
  $d.f.\{W\} = n_\bullet - I$
  $MSW = \frac{SSW}{d.f.\{W\}}$

- $SSTO = SSB + SSW$
  $d.f.\{TO\} = n_\bullet - 1$
  $MSTO = \frac{SSTO}{d.f.\{TO\}}$

Hypothesis Test ANOVA

- Step 1: State the null and alternative.

- Step 2: The test-statistic is: $F_S = \frac{MSB}{MSW}$
  with $d.f.\{numerator\} = I - 1$
  $d.f.\{denomenator\} = n_\bullet - I$

- Step 3: The p-value is:
  $Pr\{F > F_S\}$

Confidence Intervals for ANOVA

- A (1-α)100% simultaneous/overall/family-wise CI for k pairs of means is:
  $\bar{y}_a - \bar{y}_b \pm t_{\alpha/(2k)} \sqrt{MSW(\frac{1}{n_a} + \frac{1}{n_b})}$
  at d.f $= n_\bullet - I$

Confidence Intervals for a Proportion

- A $(1-\alpha)$100% confidence interval for p (the true proportion) is:
  $\tilde{p} \pm Z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$
  where $\tilde{p} = \frac{y+2}{n+4}$ and you may find
  $Z_{\alpha/2}$ with $t_{\alpha/2}$ at d.f. $= \infty$

χ^2 Goodness of Fit Test

- Step 1: State the null and alternative.

- Step 2: $e_i = np_i$,
  $\chi_S^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$ with d.f. $= k - 1$.

- Step 3: p-value $= Pr\{\chi^2 > \chi_S^2\}$

χ^2 Independence Test

- Let there be a categorical variable A with I categories, and let there be a categorical variable B with J categories.
  Step 1: State the null and alternative.

- Step 2: $e_{ij} = \frac{r_i c_j}{n}$, where r_i = row total for row i, c_j = column total for column j
  $\chi_S^2 = \sum_{alli,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ with d.f. $= (I-1)(J-1)$

- Step 3: p-value $= Pr\{\chi^2 > \chi_S^2\}$

Confidence Intervals for p_1 - p_2

- A $(1-\alpha)$100% confidence interval for $p_1 - p_2$ (a difference in true probabilities/proportions) is:
  $\tilde{p}_1 - \tilde{p}_2 \pm Z_{\alpha/2} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$
  where $\tilde{p}_1 = \frac{y_1+1}{n_1+2}$, $\tilde{p}_2 = \frac{y_2+1}{n_2+2}$ and you may find
  $Z_{\alpha/2}$ with $t_{\alpha/2}$ at d.f. $= \infty$

Linear Regression

- The estimated slope is: $b_1 = r\frac{s_y}{s_x}$

- The estimated intercept is: $b_0 = \bar{y} - b_1\bar{x}$

- The estimated error is: $e_i = y_i - \hat{y}_i$,
  where $\hat{y}_i$ is the estimated value of y based on our regression line.

- $s_e = \sqrt{\frac{SSE}{n-2}}$

- A $(1-\alpha)$100% confidence interval for the slope is:
  $b_1 \pm t_{\alpha/2} \frac{s_e}{\sqrt{(n-1)s_X^2}}$ with d.f. $= n - 2$

- If the assumptions of linear regression hold,
  $Y \sim N(\beta_0 + \beta_1 X, \sigma_\epsilon^2)$ for any value of X.

- The coefficient of determination is: r^2 (the correlation coefficient, squared)