Summary of Topics for Midterm Exam #1

STA 371G, Spring 2016

Listed below are the major topics covered in class that are likely to be in Midterm Exam #1:

- Mean (expectation), variance and standard deviation of a random variable.

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i P(X = x_i), \quad \text{Var}[X] = \sum_{i=1}^{n} (x_i - \mathbb{E}[X])^2 P(X = x_i), \quad \text{sd}[X] = \sqrt{\text{Var}[X]}$$

- Add a constant to a random variable, multiply a random variable by a constant.

  If $Y = a + bX$, then

$$\mathbb{E}[Y] = a + b\mathbb{E}[X], \quad \text{Var}[Y] = b^2\text{Var}[X], \quad \text{sd}[Y] = |\,b\,| \times \text{sd}[X].$$

- Conditional, joint and marginal probabilities.

$$P(Y = y \,|\, X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

$$P(Y = y, X = x) = P(Y = y \,|\, X = x)P(X = x)$$

$$P(Y = y) = \sum_{x} P(Y = y, X = x)$$

- Independent random variables, sum of independent random variables.
  - Two random variables X and Y are independent if $P(Y = y \,|\, X = x) = P(Y = y)$ for all possible x and y.
  - If X and Y are independent , then $P(Y = y, X = x) = P(Y = y)P(X = x)$.
  - If $Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$, then

$$\mathbb{E}[Y] = a_0 + a_1\mathbb{E}[X_1] + a_2\mathbb{E}[X_2] + \cdots + a_n\mathbb{E}[X_n].$$

    If X_i and X_j are independent for $i \neq j$, then we further have

$$\text{Var}[Y] = a_1^2\text{Var}[X_1] + a_2^2\text{Var}[X_2] + \cdots + a_n^2\text{Var}[X_n].$$

- Normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ is the mean, σ^2 is the variance, and σ is the standard deviation.
  - Probability density function: area under the curve represents probability.
  - Standard normal distribution $Z \sim \mathcal{N}(0, 1)$.
  - Standardizing a normal random variable $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
  - $P(X < x) = P(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = P(Z < \frac{x-\mu}{\sigma})$.

- $P(-2 < Z < 2) \approx 0.95$; $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$.

- Estimate μ and σ^2 when $X \sim \mathcal{N}(\mu, \sigma^2)$.
  - Use the sample mean $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ to estimate μ.
  - Use the sample variance $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ to estimate σ^2.

- Sampling distribution of a sample mean $\bar{X}$:
  - $\bar{X} \sim \mathcal{N}(\mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n})$.
  - The sampling distribution of $\bar{X}$ is useful in determining the quality of $\bar{X}$ as an estimator for the population mean μ.
  - As the population variance σ^2 is usually unknown, we use the sample variance $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ to estimate σ^2 and hence s^2/n to estimate $\sigma_{\bar{X}}^2$.
  - 95% confidence interval of μ (approximately): $\bar{X} \pm 2\sqrt{\frac{s^2}{n}}$.

- Binomial distribution and its normal approximation
  - $X \sim \text{Binomial}(n, p)$ can be approximated with $X \sim \mathcal{N}(np, np(1-p))$ if n is large enough and p is not too close to 0 or 1.
  - Estimate the population proportion p when $X \sim \text{Binomial}(n, p)$, where n is the sample size.
    * Use the sample proportion $\hat{p} = \frac{X}{n}$ to estimate p.
    * Approximately, we have $\hat{p} \sim \mathcal{N}(p, \frac{\hat{p}(1-\hat{p})}{n})$.
    * 95% confidence interval of p: $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

- Simple Linear Regression
  - Least squares estimation: given n observations $(x_1, y_1), \cdots, (x_n, y_n)$, we estimate the intercept b_0 and slope b_1 by finding a straight line $\hat{y}_i = b_0 + b_1 x_i$ that minimizes

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left[y_i - (b_0 + b_1 x_i)\right]^2.$$

  - Sample means of X and Y

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}.$$

  - Sample covariance

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

– Sample correlation

$$r_{xy} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\mathrm{Cov}(X, Y)}{s_x s_y}.$$

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}, \quad s_x = \sqrt{s_x^2}$$

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}, \quad s_y = \sqrt{s_y^2}$$

– Least squares estimation:

$$b_0 = \bar{y} - b_1 \bar{x}, \qquad b_1 = r_{xy} \frac{s_y}{s_x}$$

– Interpreting covariance, correlation and regression coefficients.

– SST, SSR, SSE

$$\mathrm{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\mathrm{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \left[(b_0 + b_1 x_i) - \bar{y}\right]^2$$

$$\mathrm{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left[y_i - (b_0 + b_1 x_i)\right]^2$$

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

Note that $\bar{y} = b_0 + b_1 \bar{x}$, $\hat{y}_i = b_0 + b_1 x_i$, $\bar{\hat{y}} = \bar{y}$, and $e_i = y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$.

– Coefficient of determination:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = r_{xy}^2$$

– Regression assumptions and statistical model.

$$Y = \beta_0 + \beta_1 X + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Assuming β_0, β_1 and σ^2 are known, given x_i, the 95% prediction interval of y_i is

$$(\beta_0 + \beta_1 x_i) \pm 2\sigma.$$