Midterm 1 Review

Terminology and Definations

- Population
- Sample
- Statistics
- Parameter
- Estimator, Estimate
- Support, Parameter Space
- Independence
- Expectation, Variance
- Moment Generating Function: $M_X(t) = E(e^{tX})$.
- Unbiasedness, Bias
- Likelihood
- Pivotal Quantity
- Confidence Interval

Chap 5.1 and Chap 5.2 Change of Variables

- Results should include both p.d.f. / p.m.f. and the support of transformed variable.
- Distinguish 1-1 mapping v.s. non 1-1 mapping.
- Use short cut only for the parts that involve 1-1 mapping.

Chap 5.3 Expectation and Variance

1. If a and b are constants,

   $E(aX + b) = aE(X) + b$ $\qquad\qquad$ $Var(aX + b) = a^2Var(X)$

2. For any transformation $u(X)$,

   $Var(u(X)) = E[u(X)^2] - E[u(X)]^2$

3. For transformations $u_1(X_1)$, $u_2(X_2)$, ... $u_n(X_n)$,

   $$E[u_1(X_1) + u_2(X_2) + ... + u_n(X_n)] = E[u_1(X_1)] + E[u_2(X_2)] + ... + E[u_n(X_n)]$$

4. If X_1, X_2, ... X_n are independent,

   $$E[u_1(X_1) \times u_2(X_2) \times ... \times u_n(X_n)] = E[u_1(X_1)] \times E[u_2(X_2)] \times ... \times E[u_n(X_n)]$$

5. If X_1, X_2, ... X_n are independent,

   $$Var[u_1(X_1) + u_2(X_2) + ... + u_n(X_n)] = Var[u_1(X_1)] + Var[u_2(X_2)] + ... + Var[u_n(X_n)]$$

Chap 5.4 The Moment Generating Function

If X_1, X_2, ... X_n are independent random variables with m.g.f. $M_{X_i}(t) = E(e^{X_i t})$, then $Y = \sum_{i=1}^n a_i X_i$ has m.g.f. $M_Y(t) = \prod_{i=1}^n E(e^{a_i X_i t})$.

Chap 5.5 Random Variables related with Normal distributions

Theorem 5.5-1: If X_1, X_2, ... X_n are independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$, then $Y = \sum_{i=1}^n c_i X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

If X_1, X_2, ... X_n are independent random variables with $X_i \sim N(\mu, \sigma^2)$

Corollary 5.5-1: $\bar{X} \sim N(\mu, \sigma^2/n)$.

Theorem 5.5-2: $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ is the sample variance,

$$\frac{S^2(n-1)}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

Theorem 5.5-3: Student's t distribution $T = \frac{Z}{\sqrt{U/r}} \sim t(r)$, where $Z \sim N(0,1)$ and $U \sim \chi^2(r)$.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Chap 5.6 The Central Limit Theorem (CLT)

With sufficiently many i.i.d. samples collected, the sample mean $\bar{X}$ follows $N(\mu, \sigma^2/n)$ approximately, regardless the true distribution of X_i.

Chap 6.4 Point Estimation

The Method of Moments (MoM)

1. Find the moments, e.g. $E(X)$, $E(X^2)$, etc.
2. Set the equations for k = 1, 2, ...

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k = E(X^k)$$

   The number of moment-based equations is the number of unknown parameters

3. Solve the equations.

The Maximum Likelihood Estimation (MLE)

1. Find the log likelihood. Note that it shall include $(x_1, x_2, \ldots, x_n)$ and the parameter of interest.
2. Find the first derivative of the log likelihood with respect to the parameters of interest, and set them to zeros.
3. Solve the equations.

Chap 7.1, 7.2, 7.3 Confidence Interval

Confidence Intervals

Parameter	Assumptions	Endpoints
μ	$N(\mu, \sigma^2)$ or n large, σ^2 known	$\overline{x} \pm z_{\alpha/2}\, \dfrac{\sigma}{\sqrt{n}}$
μ	$N(\mu, \sigma^2)$ σ^2 unknown	$\overline{x} \pm t_{\alpha/2}(n-1)\dfrac{s}{\sqrt{n}}$
$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 known	$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}$
$\mu_X - \mu_Y$	Variances unknown, large samples	$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\dfrac{s_x^2}{n} + \dfrac{s_y^2}{m}}$
$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$, unknown	$\overline{x} - \overline{y} \pm t_{\alpha/2}(n+m-2)s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}$, $s_p = \sqrt{\dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$\mu_D = \mu_X - \mu_Y$	X and Y normal, but dependent	$\overline{d} \pm t_{\alpha/2}(n-1)\dfrac{s_d}{\sqrt{n}}$
p	$b(n, p)$ n is large	$\dfrac{y}{n} \pm z_{\alpha/2}\sqrt{\dfrac{(y/n)[1-(y/n)]}{n}}$
$p_1 - p_2$	$b(n_1, p_1)$ $b(n_2, p_2)$	$\dfrac{y_1}{n_1} - \dfrac{y_2}{n_2} \pm z_{\alpha/2}\sqrt{\dfrac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \dfrac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$, $\widehat{p_1} = y_1/n_1,\ \widehat{p_2} = y_2/n_2$

Midterm 2 Review

Terminology and Definations

- Deterministic relationship v.s. statistical relationship.
- Interpretations of α and β in linear regression.
- What is the key difference between Least Square Estimation and Maximum Likelihood Estimation for linear regression in term of model assumptions.
- Confidence interval v.s. Prediction interval.
- Factors that affect the width of those intervals.
- Null v.s. Alternative hypothesis.
- Type I error v.s. Type II error.

Chap 6.5 and 7.6 Linear Regression

$\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of random variables, Y_i's.

$$\hat{\alpha} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\sigma}^2 = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2$$

$(1-a) \times 100\%$ CI for α is

$$\hat{\alpha} \pm t_{\frac{a}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{(n-2)}}.$$

$(1-a) \times 100\%$ CI for β is

$$\hat{\beta} \pm t_{\frac{a}{2},n-2}\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Confidence interval for $E(Y_i|x_i)$ is

$$\hat{\alpha} + \hat{\beta}(x_i - \bar{x}) \pm t_{\frac{a}{2},n-2}\sqrt{\frac{n\hat{\sigma}^2}{n-2} \times (\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}.$$

Prediction interval for $Y_{n+1}|x_i$ is

$$\hat{\alpha} + \hat{\beta}(x_i - \bar{x}) \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{n\hat{\sigma}^2}{n-2} \times (1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}.$$

Chap 7.4 Sample Size Calculation

The sample size necessary for estimating a population mean μ with $100(1-\alpha)\%$ confidence and error no larger than ϵ is:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{\epsilon^2}.$$

When σ^2 is unknown, we replace σ^2 by sample variance S^2.

The sample size necessary for estimating a population proportion p with $100(1-\alpha)\%$ confidence and error no larger than ϵ is:

1. Guess the value of p, say p^* based on prior knowledge or use a pilot study to find p^*,

$$n = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\epsilon^2}.$$

2. We know that when $p = 0.5$, the value of $p(1-p)$ is maximized,

$$n = \frac{z_{\alpha/2}^2 0.5(1-0.5)}{\epsilon^2} = \frac{z_{\alpha/2}^2}{4\epsilon^2}.$$

Since sample size n needs to be an integer, we round it up.

Chap 8.1, 8.2, 8.3 Hypothesis Test

Tests of Hypotheses

Hypotheses	Assumptions	Critical Region
H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$	$N(\mu, \sigma^2)$ or n large, σ^2 known	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \ge z_\alpha$
H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$	$N(\mu, \sigma^2)$ σ^2 unknown	$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} \ge t_\alpha(n-1)$
H_0: $\mu_X - \mu_Y = 0$ H_1: $\mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 known	$z = \dfrac{\overline{x} - \overline{y} - 0}{\sqrt{(\sigma_X^2/n) + (\sigma_Y^2/m)}} \ge z_\alpha$
H_0: $\mu_X - \mu_Y = 0$ H_1: $\mu_X - \mu_Y > 0$	Variances unknown, large samples	$z = \dfrac{\overline{x} - \overline{y} - 0}{\sqrt{(s_x^2/n) + (s_y^2/m)}} \ge z_\alpha$
H_0: $\mu_X - \mu_Y = 0$ H_1: $\mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$, unknown	$t = \dfrac{\overline{x} - \overline{y} - 0}{s_p\sqrt{(1/n) + (1/m)}} \ge t_\alpha(n+m-2)$ $s_p = \sqrt{\dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
H_0: $\mu_D = \mu_X - \mu_Y = 0$ H_1: $\mu_D = \mu_X - \mu_Y > 0$	X and Y normal, but dependent	$t = \dfrac{\overline{d} - 0}{s_d/\sqrt{n}} \ge t_\alpha(n-1)$
H_0: $p = p_0$ H_1: $p > p_0$	$b(n, p)$ n is large	$z = \dfrac{(y/n) - p_0}{\sqrt{p_0(1-p_0)/n}} \ge z_\alpha$
H_0: $p_1 - p_2 = 0$ H_1: $p_1 - p_2 > 0$	$b(n_1, p_1)$ $b(n_2, p_2)$	$z = \dfrac{(y_1/n_1) - (y_2/n_2) - 0}{\sqrt{\left(\dfrac{y_1+y_2}{n_1+n_2}\right)\left(1 - \dfrac{y_1+y_2}{n_1+n_2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \ge z_\alpha$