Concept List

Chapter 9: Two-Sample Inference

Terms:
1. Two-sample data
2. $T-$distribution
3. $F-$distribution
4. $\chi^2$ distribution
5. Two-sample $t-$test
6. Pooled sample variance
7. Behrens-Fisher problem

Formulas:

1. $\chi^2(n) + \chi^2(m) = \chi^2(n + m)$
2. A students-$T$ with $n$ degrees of freedom is the quotient of a standard normal with an independent $\chi^2(n)$, i.e.
   $$z \sqrt{\frac{\chi^2(n)}{n}}$$
3. The pooled variance is
   $$S_p^2 = \frac{(n - 1)S_X^2 + (m - 1)S_Y^2}{n + m - 2}$$
4. If $X_1, \ldots, X_n$ are iid samples from a $N(\mu_x, \sigma)$ distribution and $Y_1, \ldots, Y_m$ are iid samples from a $N(\mu_y, \sigma)$ distribution, then
   $$T = \frac{X - Y - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$
   is distributed students-$T$ with $n + m - 2$ degrees of freedom. (Here, $S_p$ is the pooled variance.)
5. **Hypothesis Testing** ($H_0 : \mu_x = \mu_y$.) Suppose $X_1, \ldots, X_n$ are iid samples from a $N(\mu_x, \sigma)$ distribution and $Y_1, \ldots, Y_m$ are iid samples from a $N(\mu_y, \sigma)$ distribution. To test a null hypothesis of the form $H_0 : \mu_x = \mu_y$, form the test statistic $t$ as in (1). The critical region depends on both the level of significance $\alpha$ and the alternative hypothesis $H_1$, as follows:
   
   (a) If $H_1 : \mu_x > \mu_y$, then the critical region is $t > t_{\alpha,n+m-2}$.
   (b) If $H_1 : \mu_x < \mu_y$, then the critical region is $t < t_{1-\alpha,n+m-2}$.
   (c) If $H_1 : \mu_x \neq \mu_y$, then the critical region is $t > t_{\alpha/2,n+m-2}$ or $t < t_{1-\alpha/2,n+m-2}$.
6. **Confidence Intervals** (for $\mu_x - \mu_y$.) Suppose $X_1, \ldots, X_n$ are iid samples from a $N(\mu_x, \sigma)$ distribution and $Y_1, \ldots, Y_m$ are iid samples from a $N(\mu_y, \sigma)$ distribution. A 100$(1 - \alpha)$% confidence interval for $\mu_x - \mu_y$ is
   $$[\overline{x} - \overline{y} - s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \cdot t_{\alpha/2,m+n-2}, \overline{x} - \overline{y} + s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \cdot t_{\alpha/2,m+n-2}]$$
7. If $X_1, \cdots, X_n$ are iid samples from a $N(\mu_x, \sigma)$ distribution and $Y_1, \cdots, Y_m$ are iid samples from a $N(\mu_y, \sigma)$ distribution, then
\[
s = \frac{S_Y^2}{S_X^2} \sim F(m-1, n-1). \tag{2}
\]

8. **Hypothesis Testing ($H_0 : \sigma_x = \sigma_y$)** Suppose $X_1, \cdots, X_n$ are iid samples from a $N(\mu_x, \sigma_x)$ distribution and $Y_1, \cdots, Y_m$ are iid samples from a $N(\mu_y, \sigma_y)$ distribution. To test a hypothesis of the form $H_0 : \sigma_x = \sigma_y$, form a test statistic of the form (2), above. The critical region depends on both the significance level $\alpha$ and the alternative hypothesis $H_1$, as follows:

(a) If $H_1 : \sigma_x^2 > \sigma_y^2$, then the critical region is $s < F_{\alpha,m-1,n-1}$.
(b) If $H_1 : \sigma_x^2 < \sigma_y^2$, then the critical region is $s > F_{1-\alpha,m-1,n-1}$.
(c) If $H_1 : \sigma_x^2 \neq \sigma_y^2$, then the critical region is $s < F_{\alpha/2,m-1,n-1}$ or $s > F_{1-\alpha/2,m-1,n-1}$.

9. **Confidence Intervals (for $\sigma^2_x/\sigma^2_y$)** Suppose $X_1, \cdots, X_n$ are iid samples from a $N(\mu_x, \sigma)$ distribution and $Y_1, \cdots, Y_m$ are iid samples from a $N(\mu_y, \sigma)$ distribution. A $100(1-\alpha)%$ confidence interval for $\sigma^2_y/\sigma^2_x$ is
\[
\left[ \frac{s_Y^2}{s_X^2} F_{\alpha/2,m-1,n-1}, \frac{s_Y^2}{s_X^2} F_{1-\alpha/2,m-1,n-1} \right].
\]

10. **(Binomial data)** Suppose $X_1, \cdots, X_n$ are samples from a binary distribution with parameter $p_x$, and $Y_1, \cdots, Y_m$ are samples from a binary distribution with parameter $p_y$. Then
\[
z = \frac{x/n - \mu}{\sqrt{p_x(1-p_x)/n + p_y(1-p_y)/m}}
\]
is distributed approximately normally.

11. **Hypothesis Testing (binomial data)** Suppose $X_1, \cdots, X_n$ are samples from a binary distribution with parameter $p_x$, and $Y_1, \cdots, Y_m$ are samples from a binary distribution with parameter $p_y$. To test $H_0 : p_x = p_y$ at the $\alpha$ level, form
\[
z = \frac{x/n - \mu}{\sqrt{p_x(1-p_x)/n + p_y(1-p_y)/m}},
\]
where
\[p_e = \frac{x+y}{n+m}.
\]
Then the critical region depends on the alternative hypothesis $H_1$ as follows:

(a) If $H_1 : p_x > p_y$, reject $H_0$ if $z > z_{\alpha}$.
(b) If $H_1 : p_x > p_y$, reject $H_0$ if $z < -z_{\alpha}$.
(c) If $H_1 : p_x \neq p_y$, reject $H_0$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$.