# TDS2201 Statistical Data Analysis

taken in Trimester 1, 2017/2018

Summary Sheet by Kia Kin

## Univariate Data Exploration

1. Five-number summary: Min, $Q_1$, Median, $Q_3$, Max
2. $IQR = Q_3 - Q_1$
   lower boundary $= Q_1 - 1.5 \times IQR$
   outlier < lower boundary
   upper boundary $= Q_3 + 1.5 \times IQR$
   outlier > upper boundary

## Sampling Distribution

1. To find $P(X)$, use $Z = \dfrac{x-\mu}{\sigma}$
   To find $P(\bar{X})$, use $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
2. $P(Z > a) = 1 - P(Z < a)$
   $P(a < Z < b) = P(Z < b) - P(Z < a)$

## Confidence Interval $100(1-\alpha)\%$

1. CI for $\mu$
   (a) $\bar{x} \pm Z_{\frac{\sigma}{2}} \dfrac{\sigma}{\sqrt{n}}$
   Use when 1. $\sigma$ known. 2. $\sigma$ unknown. $n \geq 30$. (Swap $\sigma$ with $s$.)
   (b) $\bar{x} \pm t_{\frac{\sigma}{2},v} \dfrac{s}{\sqrt{n}}$
   Use when $\sigma$ unknown. $n < 30$.
2. CI for $\mu_1 - \mu_2$
   (a) $(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\sigma}{2}} \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
   Use when 1. $\sigma_1^2, \sigma_2^2$ known. 2. $\sigma_1^2$, $\sigma_2^2$ unknown. $n_1 \geq 30$. $n_2 \geq 30$. (Swap $\sigma$ with $s$.)
   (b) $(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\sigma}{2},n_1+n_2-2} S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$
   where $S_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
   Use when $\sigma_1^2$, $\sigma_2^2$ unknown. $n_1 < 30$. $n_2 < 30$. $\sigma_1^2 = \sigma_2^2$.
   (c) $(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\sigma}{2},v} \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
   where $w_1 = \dfrac{s_1^2}{n_1}$, $w_2 = \dfrac{s_2^2}{n_2}$, $v = \dfrac{(w_1+w_2)^2}{\frac{w_1^2}{n_1-1} + \frac{w_2^2}{n_2-1}}$
   Use when $\sigma_1^2$, $\sigma_2^2$ unknown. $n_1 < 30$. $n_2 < 30$. $\sigma_1^2 \neq \sigma_2^2$.
3. CI for $\mu_d$ (paired data)
   $\bar{d} \pm t_{\frac{\sigma}{2},v} \dfrac{s_d}{\sqrt{n}}$
4. CI for $p$
   $\hat{p} \pm Z_{\frac{\sigma}{2}} \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$
   where $\hat{p} = \dfrac{X}{N}$, $\hat{q} = 1 - \hat{p}$
   Use when $n\hat{p} \geq 5$. $n\hat{q} \geq 5$.
5. CI for $p_1 - p_2$
   $(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\sigma}{2}} \sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}$
6. CI for $\sigma^2$
   $\dfrac{(n-1)s^2}{\chi^2_{\frac{\sigma}{2},v}} < \sigma^2 < \dfrac{(n-1)s^2}{\chi^2_{1-\frac{\sigma}{2},v}}$
7. CI for $\dfrac{\sigma_1^2}{\sigma_2^2}$
   $\dfrac{s_1^2}{s_2^2} \cdot \dfrac{1}{f_{\frac{\sigma}{2},v_1,v_2}} < \dfrac{\sigma_1^2}{\sigma_2^2} < \dfrac{s_1^2}{s_2^2} \cdot f_{\frac{\sigma}{2},v_2,v_1}$
8. To find appropriate sample size, use margin of error $= \varepsilon$

## Test of Significance

1. (a) State $H_0$, $H_1$
   (b) test statistic
   (c) p-value
   (d) p-value $\leq \alpha \Rightarrow$ reject $H_0$
   p-value $> \alpha \Rightarrow$ do not reject $H_0$
   $\alpha$ is level of significance
   (e) Conclusion by restating $H_1$
2. Direction of $H_1$
   (a) >, < - same
   (b) $\geq$, $\leq$ - flip
   (c) unequal $\neq$
3. Test for $\mu$
   (a) $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
   Use when 1. $\sigma$ known. 2. $\sigma$ unknown. $n \geq 30$. (Swap $\sigma$ with $s$.)

(b) $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$
Use when $\sigma$ unknown. $n < 30$.

4. Test for $\mu_1 - \mu_2$

(a) $Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Use when 1. $\sigma_1^2$, $\sigma_2^2$ known. 2. $\sigma_1^2, \sigma_2^2$ unknown. $n_1 \geq 30$. $n_2 \geq 30$. (Swap $\sigma$ with $s$.)

(b) $T = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
where $S_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
Use when $\sigma_1^2$, $\sigma_2^2$ unknown. $n_1 < 30$. $n_2 < 30$. $\sigma_1^2 = \sigma_2^2$.

(c) $T = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
where $w_1 = \dfrac{s_1^2}{n_1}$, $w_2 = \dfrac{s_2^2}{n_2}$, $v = \dfrac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}}$
Use when $\sigma_1^2$, $\sigma_2^2$ unknown. $n_1 < 30$. $n_2 < 30$. $\sigma_1^2 \neq \sigma_2^2$.

5. Test of $\mu_d$
$T = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$

6. Test for $p$
$Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$ where $q_0 = 1 - p_0$

7. Test for $p_1 - p_2$
$Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$

8. Test for $\sigma^2$
$\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$

9. Test for $\dfrac{\sigma_1^2}{\sigma_2^2}$
$F = \dfrac{s_1^2}{s_2^2} \cdot \dfrac{\sigma_2^2}{\sigma_1^2}$

10. p-value

(a) $H_1 : < \ \Rightarrow P(<)$
(b) $H_1 : > \ \Rightarrow P(>)$
(c) $H_1 : \neq \ \Rightarrow 2 \times P(>)$
(d) no exact value
Example. p-value $= P(t > 1.989, v = 18)$
$\alpha = 0.05 \Rightarrow T_{0.05,18} = 1.734$
$T_{0.025,18} = 2.101$
$0.025 <$ p-value $< 0.05$

11. (a) False Positive,
$P(\text{Type I Error}) = P(\text{Reject } H_0 \mid H_0 \text{ is True})$
(b) False Negative,
$P(\text{Type II Error}) = P(\text{Not reject } H_0 \mid H_0 \text{ is False})$
(c) Power $= 1 - P(\text{Type II Error})$

# Test of Independence

1. $H_0$ : 2 categorical variables are not associated / dependent.

   $H_1$ : 2 categorical variables are associated / dependent.

2. expected count $= \dfrac{\text{row total} \times \text{column total}}{\text{table total}}$

   Example.

| | Drug X | Placebo | |
|---|---|---|---|
| Insonmia | 91 (69) | 13 (35) | 104 |
| No insonmia | 271 (293) | 170 (148) | 441 |
| | 362 | 183 | 545 |

$\dfrac{362 \cdot 104}{545} = 69$

3. test statistic $\chi^2 = \Sigma\Sigma\dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$

   $v = (r-1)(c-1)$

   $O$ is Observed, $E$ is Expected.

# Linear Regression

1. $y = \beta_0 + \beta_1 x + \varepsilon$
   Fitted regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

2. Residual = observed $y$ − predicted $\hat{y}$

3. CI for $\beta_i$
   $\hat{\beta}_i \pm t_{\frac{\alpha}{2}, n-2} se(\hat{\beta}_i)$

4. Test for $\beta_i$

   (a) $H_0 : \beta_i = b$ vs. $H_1 : \beta_i \neq b$
   (b) test statistic $T = \dfrac{\hat{\beta}_i - b}{se(\hat{\beta}_i)}$, $v = n - 2$
   (c) Special case: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

5. CI for $\hat{y}_0$

   $$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(x_0 - \bar{x})}{S_{xx}})}$$

   where $S_{xx} \approx (\frac{\hat{\sigma}}{se(\beta_1)})^2$

6. PI for $\hat{y}_0$

   $$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})}{S_{xx}})}$$

   where $S_{xx} \approx (\frac{\hat{\sigma}}{se(\beta_1)})^2$

7. `R` output

   (a) `Estimate` - $\beta_0$, $\beta_1$

   (b) `Std.Error` - $se(\beta_0)$, $se(\beta_1)$

   (c) `Residual standard error` - $\hat{\sigma}$

8. Example. $r^2 = 0.9025$.

   90.25% of variation in y-variable can be explained by the model with x-variable.

# CASIO 570MS

## Mean and Standard Deviation

1. `MODE` twice

2. `1` (`SD`)

3. Type all `x` + `M+`

4. `SHIFT` + `2` (`S-VAR`)

5. `1` (Mean)

   `2` ($\sigma$)

   `3` ($s$)

## $P(Z < z)$ and $P(Z > z)$

1. `MODE` twice

2. `1` (`SD`)

3. `SHIFT` + `3` (`DISTR`)

4. `1` ($P(Z < z)$)

   `3` ($P(Z > z)$)

## Linear Regression

1. `MODE` twice

2. `REG`

3. `Lin`

4. Type all `x,y M+`

5. `SHIFT` + `2`

6. `>` twice

7. `A` $= \beta_0 =$ y-intercept

   `B` $= \beta_1 =$ slope

   `r` $=$ correlation coefficient

# Statistical Tables

1. Example. $Z_{0.025} = k$ / $P(Z < k) = 0.025$

   | $Z$ | .06 |
   |---|---|
   | $-1.9$ | .0250 |

   $k = -1.96$

2. Example. $t_{0.025,6} = k$

   | | $A$ |
   |---|---|
   | $v$ | .025 |
   | 6 | 2.447 |

   $k = 2.447$

3. Example. $\chi^2_{0.025,16} = k$

   | | $\alpha$ |
   |---|---|
   | $v$ | 0.025 |
   | 16 | 28.845 |

   $k = 28.845$

4. Example. $f_{0.01,7,3} = k$

   | | $f_{0.01(v_1,v_2)}$ |
   |---|---|
   | | $v_1$ |
   | $v_2$ | 7 |
   | 3 | 27.67 |

   $k = 27.67$