# 2019R1 Discrete Data Analysis (STAT5107) Final Test

*Yiu Chung WONG 1155017920*

```r
set.seed(5107);
```

### 1.

a) Nominal
b) Ordinal
c) Ordinal
d) Nominal
e) Nominal
f) Ordinal

### 2a.

$$\binom{n}{y}\pi^y(1-\pi)^{n-y}$$

where $n = 100$, $\pi = 0.25$, $y =$ number of correct answer, ranging from 0 to $n$

### 2b.

```r
n <- 100
prob <- 0.25
mean <- n * prob
var <- n * prob * (1-prob)
sd_away <- (50-mean)/sqrt(var)
p <- pbinom(q = 49, prob = prob, size = n, lower.tail = FALSE) #at least 50; P[X > 49]
```

- 50 or more correct responses is equivalent to 5.7735027 standard deviations or higher away from the mean. The probability of this happening is $6.6385025 \times 10^{-8}$. It's not happening.

**3a.**

**Interchanging rows with columns**

$$\begin{matrix} n_{22} & n_{21} \\ n_{12} & n_{11} \end{matrix}$$

we would have

$$\begin{matrix} \pi_{22} & \pi_{21} \\ \pi_{12} & \pi_{11} \end{matrix}$$

$$\hat{\theta} = \frac{\hat{\pi}_{22}\hat{\pi}_{11}}{\hat{\pi}_{12}\hat{\pi}_{21}} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}}$$

**3b.**

$$\begin{matrix} ckn_{11} & cn_{12} \\ kn_{21} & n_{22} \end{matrix}$$

$$\hat{\theta} = \frac{\frac{ckn_{11}}{n}\frac{n_{22}}{n}}{\frac{cn_{12}}{n}\frac{kn_{21}}{n}} = \frac{\frac{n_{11}}{n}\frac{n_{22}}{n}}{\frac{n_{12}}{n}\frac{n_{21}}{n}} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{12}\hat{\pi}_{21}}$$

**3c.**

Relative Risk: $\frac{n_{11}(n_{21}+n_{22})}{n_{21}(n_{11}+n_{12})}$

Diff of Proportion: $\frac{n_{11}}{n_{11}+n_{12}} - \frac{n_{21}}{n_{21}+n_{22}}$

**Interchanging rows with columns**

$$\begin{matrix} n_{22} & n_{21} \\ n_{12} & n_{11} \end{matrix}$$

Relative Risk is now: $\frac{n_{22}(n_{12}+n_{11})}{n_{12}(n_{22}+n_{21})}$

Diff of Proportion is now: $\frac{n_{22}}{n_{22}+n_{21}} - \frac{n_{12}}{n_{12}+n_{11}}$

**Multiply by constant**

$$\begin{matrix} ckn_{11} & cn_{12} \\ kn_{21} & n_{22} \end{matrix}$$

Relative Risk is now: $\frac{ckn_{11}(kn_{21}+n_{22})}{kn_{21}(ckn_{11}+cn_{12})}$

Diff of Proportion is now: $\frac{ckn_{11}}{ckn_{11}+cn_{12}} - \frac{kn_{21}}{kn_{21}+n_{22}}$

**3d.**

**Theorem: odds ratio equals to one i.i.f. row variable X and the column variable Y are independent**

Proof:

1) Assume odds ratio equals to one

$$\theta = 1 \implies \Omega_1 = \Omega_2 \implies \frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{21}}{\pi_{22}}$$

$$\begin{aligned}
\Pr(Y=1|X=1) &= \frac{\Pr(Y=1, X=1)}{\Pr(X=1)} = \frac{\pi_{11}}{\pi_{1+}} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}} = \frac{1}{1+\frac{\pi_{12}}{\pi_{11}}} \\
\Pr(Y=1|X=2) &= \frac{\Pr(Y=1, X=2)}{\Pr(X=2)} = \frac{\pi_{21}}{\pi_{2+}} = \frac{\pi_{21}}{\pi_{21}+\pi_{22}} = \frac{1}{1+\frac{\pi_{22}}{\pi_{21}}}
\end{aligned}$$

we have

$$\Pr(Y=1|X=1) = \Pr(Y=1|X=2) = p^*$$

Also

$$\begin{aligned}
\Pr(Y=1) = \pi_{1+} &= \pi_{11} + \pi_{12} \\
&= \pi_{1+}\Pr(Y=1|X=1) + \pi_{2+}\Pr(Y=1|X=2) \\
&= \pi_{1+}p^* + \pi_{2+}p^* \\
&= (\pi_{1+} + \pi_{2+})p^* \\
&= p^*
\end{aligned}$$

Hence

$$\begin{aligned}
\Pr(Y=1|X=1) = \Pr(Y=1|X=2) = \Pr(Y=1) \\
\Pr(Y=2|X=1) = \Pr(Y=2|X=2) = \Pr(Y=2)
\end{aligned}$$

2) Assume X and Y are independent

$$\begin{aligned}
\theta &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \\
&= \frac{\Pr(X=1, Y=1)\Pr(X=2, Y=2)}{\Pr(X=1, Y=2)\Pr(X=2, Y=1)} \\
&\overset{Ind.}{=} \frac{\Pr(X=1)\Pr(Y=1)\Pr(X=2)\Pr(Y=2)}{\Pr(X=1)\Pr(Y=2)\Pr(X=2)\Pr(Y=1)} \\
&= 1
\end{aligned}$$

### 4.

$$prob = \frac{odds}{1 + odds}$$

```
italy_odds <- 11/10
italy_prob <- italy_odds / (1 + italy_odds)

bulgaria_odds <- 3/10
bulgaria_prob <- bulgaria_odds / (1 + bulgaria_odds)
```

- The probability of Italy winning is 0.5238095; the probability of Bulgaria winning is 0.2307692.
- odds ratio of Italy winning is 3.6666667
- Relative risk of Italy winning is 2.2698413; Italy is more than twice as likely to win

### 5a.

```
crab_model <- function(c1, c2, c3, width)
{
  -12.715 + 1.330*c1 + 1.402*c2 + 1.106*c3 + 0.468*width
}
inverse_logit <- function(x) 1/(1+exp(-(x)))

medium_dark_prob <- inverse_logit(crab_model(0, 0, 1, 20))
dark_prob <- inverse_logit(crab_model(0, 0, 0, 20))

prob_ratio <- medium_dark_prob/dark_prob
```

- Ratio of probabilities is 2.8292507

### 5b.

```
medium_dark_odds <- medium_dark_prob / (1-medium_dark_prob)
dark_odds <- dark_prob / (1-dark_prob)
odds_ratio <- medium_dark_odds/dark_odds
```

- Odds for medium dark is 0.1055047
- Odds for dark is 0.0349094

### Odds

$$\begin{aligned} log\frac{\pi(x)}{1-\pi(x)} &= -12.715 + 1.330 * c1 + 1.402 * c2 + 1.106 * c3 + 0.468 * width \\ \frac{\pi(x)}{1-\pi(x)} &= \exp(-12.715) \times \exp(1.330 * c1) \times \exp(1.402 * c2) \times \exp(1.106 * c3) \times \exp(0.468 * width) \end{aligned}$$

$$\begin{aligned} Odds_{medium\ dark} &= \exp(-12.715) \times \exp(1.330 * 0) \times \exp(1.402 * 0) \times \exp(1.106 * 1) \times \exp(0.468 * 20) \\ &= \exp(-12.715) \times \exp(1.106 * 1) \times \exp(0.468 * 20) \end{aligned}$$

$$\begin{aligned} Odds_{dark} &= \exp(-12.715) \times \exp(1.330 * 0) \times \exp(1.402 * 0) \times \exp(1.106 * 0) \times \exp(0.468 * 20) \\ &= \exp(-12.715) \times \exp(0.468 * 20) \end{aligned}$$

**Odds Ratio**

$$\begin{aligned} Odds\ Ratio &= \frac{Odds_{medium\ dark}}{Odds_{dark}} \\ &= \frac{\exp(-12.715) \times \exp(1.106 * 1) \times \exp(0.468 * 20)}{\exp(-12.715) \times \exp(0.468 * 20)} \\ &= \exp(1.106) \end{aligned}$$

- Odds ratio is 3.0222452
- On average, medium dark colour has 3.0222452 times the chance of having a satellite than that of dark.

**6a.**

Model 1 can be written as

$$\Pr(Y = 1 \mid X = x) = \frac{1}{1 + e^{-X^T\beta}}$$

the logistic distribution have cdf

$$F(x) = \frac{1}{1 + e^{-\frac{x-\mu}{\sigma}}}$$

Say the outcome of interest ($Y = 1$) occurs when $\widetilde{Y}$ for some threshold $C$. Then

$$\Pr(Y = 0 \mid X = x) = \Pr(\widetilde{Y} \leq C \mid X = x) = F(C; x)$$
$$\Pr(Y = 1 \mid X = x) = \Pr(\widetilde{Y} > C \mid X = x) = 1 - F(C; x)$$

if we assume the latent variable $\widetilde{Y}$ has error $U$ that follows logistic distribution, and assuming the linear predictor $X^T\beta$ represent the mean $\mu$ of the logistic distribution: $\mu = X^T\beta$:

$$\Pr(Y = 1 \mid x) = 1 - \frac{1}{1 + e^{-\frac{C - X^T\beta}{\sigma}}} = \frac{e^{\frac{X^T\beta\ -\ C}{\sigma}}}{1 + e^{\frac{X^T\beta\ -\ C}{\sigma}}}$$

**6b.**

**Assumptions**

1. Consistency
2. Expectation of first order derivative is zero.
3. Twice differentiable at $\beta_0$
4. $\beta_0$ strictly minimises the log likelihood.
5. $x_i \sim i.i.d.l(x, \beta_0)$ for $\beta_0$ in sample space.
6. $f(x, \beta_0) \neq f(x, \beta)$ for any $\beta \neq \beta_0$ in sample space.

**Asymptotic Distribution for the Logistic Regression model**

- Asymptotic distribution in nonlinear models follows from Taylor expansion of log likelihood $\widehat{L}(\beta)$ around the limit $\beta_0$
- Can repeat argument for this case, starting with first order condition for maximum

$$-\frac{\partial}{\partial \beta}\widehat{L}(\widehat{\beta}_{MLE}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}log(l(x, \widehat{\beta})) = 0$$

- Taylor expand around $\beta_0$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}log(l(x, \beta_0)) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial \beta^2}log(l(x, \bar{\beta}))(\widehat{\beta}_{MLE} - \beta_0)$$

- Scale by $\sqrt{n}$ and rearrange to obtain

$$\sqrt{n}(\widehat{\beta}_{MLE} - \beta_0) = (\frac{-1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial \beta^2}log(l(x, \bar{\beta})))^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}log(l(x, \beta_0))$$

- The standardized first order derivative follows central limit theorem

$$\sqrt{n}\frac{\partial}{\partial \beta}\widehat{L}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}log(l(x, \beta_0)) \xrightarrow{d} N(0, \Sigma_Q)$$

$$\Sigma_Q = E[\frac{\partial}{\partial \beta}log(l(x, \beta_0))\frac{\partial}{\partial \beta}log(l(x, \beta_0))']$$

- Second order derivative converges by uniform law of large numbers

$$\frac{-\partial^2}{\partial \beta^2}\widehat{L}(\widehat{\beta}) = \frac{-1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial \beta^2}log(l(x, \bar{\beta})) \xrightarrow{p} \mathcal{J} := -E[\frac{\partial^2}{\partial \beta^2}log(l(x, \beta_0))]$$

- The asymptotic distribution is then

$$\sqrt{n}(\widehat{\beta}_{MLE} - \beta_0) \xrightarrow{d} N(0, \mathcal{J}^{-1}\Sigma_Q\mathcal{J}^{-1})$$