ST2137 Cheatsheet AY24/25 —— @Jin Hang

Exploring Quantitative Data

Numerical Summary

- $\bar{X}$ and $X_{1/2}$ are close to each other $\Rightarrow$ symmetric.
- Mean is sensitive outliers but the **Median is not**
- $\bar{X} \gg X_{1/2} \Rightarrow$ Right Skewed
- $\bar{X} \ll X_{1/2} \Rightarrow$ Left Skewed

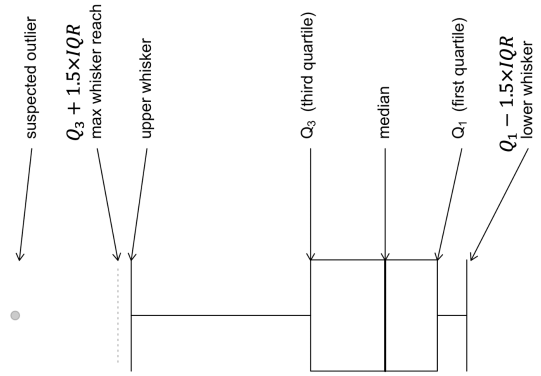
Histogram

- What is the overall pattern cluster together/gap
- Single mound or peak? unimodal/bimodal/multimodal
- Distribution symmetric or skewed
- Any suspected outliers?

Density Plot $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$

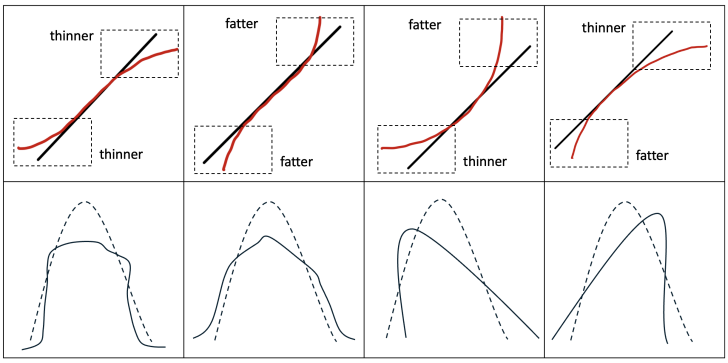
- K is a density function. A typical choice is the standard normal. The kernel places greater weights on nearby points (to x)
- h is a bandwidth, which determines which of the nearest points are used. The effect is similar to the number of bins in a histogram.
- Density Plots could overlay data for **closer comparison**, while histogram can not!

Box Plot



QQ Plot

- Plots the standardized sample quantiles against the theoretical quantiles of a N(0,1) distribution
- Points fall on a straight line $\Rightarrow$ data came from a normal distribution
- Esp. for unimodal datasets, points in the middle will typically fall close to the line.



Correlation

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \in [-1,1]$$

Exploring Categorical Data

χ^2-Test For Independence

H_0 : The two variables are indepdendent.

H_1 : The two variables are not indepdendent.

Set Significance level 5%

- If $p-\text{value} < 0.05$, Reject H_0, The two variables are not independent.
- If $p-\text{value} > 0.05$, not enough evidence to reject H_0, The two variables are independent.

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{\text{Total sample size}}$$

$$\chi^2 = \sum \frac{(|\text{expected} - \text{observed}| - 0.50)^2}{\text{expected count}}$$

χ^2-Test for $r \times c$ Tables

- Under the null hypothesis, the test statistic follows a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Fisher's Exact Test

Let W follows a hypergeometric distribution. The pmf is:

$$P(W=w) = \frac{\binom{n}{w}\binom{m}{k-w}}{\binom{n+m}{k}}$$
$$p-\text{value} = P(W \le w)$$

Odds Ratio

Measures of Association

$$OR = \frac{\text{Odds success of X}}{\text{Odds success of Y}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \in (0,\infty)$$

- OR=1 $\Rightarrow$ X,Y are independent
- Deviations from 1 indicate stronger association between the variables.
- Association OR and 1/OR are equivalent

$$\log OR = \log \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \in (-\infty,\infty)$$

Build confidence interval

1. The sample data in a 2x2 table can be labelled as $n_{11}, n_{12}, n_{21}, n_{22}$.
2. The sample odds ratio is $\widehat{OR} = \frac{n_{11}\times n_{22}}{n_{12}\times n_{21}}$
3. For a large sample size, it can be shown that $\log\widehat{OR}$ follows a Normal distribution. Hence a 95%-CI
$$\log\frac{n_{11}\times n_{22}}{n_{12}\times n_{21}} \pm z_{0.025}\times ASE(\log\widehat{OR})$$
where ASE (Asymptotic Standard Error) of the estimator is
$$\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}$$

Ordinal Variables Association

- **Concordant Pair:** Subject ranked higher on both X and Y.
- **Discordant Pair:** Subject ranking higher on X ranks lower on Y.
- **Tied Pair:** Subjects have the same classification on X and/or Y.

$$\text{Goodman-Kruskal } \gamma = \frac{C-D}{C+D}$$

$$\text{Kendall } \tau_b = \frac{C-D}{A}$$

- values close to 0 indicate a very weak trend
- values close to 1 (or -1) indicate a strong positive (negative) association

Robust Statistics

Assessing Robustness

Asymptotic Relative Efficiency (ARE): Relative efficiency of $\tilde{\theta}$ relative to $\hat{\theta}$

$$ARE(\tilde{\theta};\hat{\theta}) = \lim_{n\to\infty}\frac{\text{variance of }\hat{\theta}}{\text{variance of }\tilde{\theta}}$$

- $\hat{\theta}$ is the optimal estimator
- When using $\hat{\theta}$, we only need ARE times as many observations as when using $\tilde{\theta}$
- $\downarrow$ ARE $\Rightarrow$ $\hat{\theta}$ is better than $\tilde{\theta}$
- The sample median is less efficient than the sample mean, when the true distribution is **Normal**.
$$ARE(X_{(1/2)};\bar{X}) = 2/\pi \approx 64\%$$
- When the underlying distribution is **Normal**
$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$
$\tilde{\sigma}^2 = d^2\pi/2$, where $d = \frac{1}{n}\sum_i |Y_i - \tilde{Y}|$
$$ARE(\tilde{\sigma}^2;\hat{\sigma}^2) = 87.6\%$$

Requirements of Robust Summaries

- **Qualitative Robustness:** If the distribution F changes slightly, then the estimate should not change too much.
- **Infinitesimal Robustness:** Influence function reflects the influence of adding one more observation to a large sample.
- **Quantitative Robustness:** Consider
$$F_{x,\epsilon} = (1-\epsilon)F + \epsilon\Delta_x$$
  - Δ_x is degenerate probability distribution at x
  - **Breakdown Point:** $\epsilon_{\min} \uparrow \to \infty$ as $x \uparrow$
  - Sample Mean breakdown point: $\epsilon = 0$
  - Sample Median breakdown point: $\epsilon = 0.5$

Measures of Location

1. **Trimmed Mean**
$$\mu_t = \int_{q_{f,\gamma}}^{q_{f,1-\gamma}} x\frac{f(x)}{1-2\gamma}dx \quad \hat{\mu_t} = X_t = \frac{X_{(g+1)} + \cdots X_{(n-g)}}{n-2g}$$
   - Truncated function has to be renormalised to be a pdf.
   - Recommended value of γ is (0, 0.2]
   - Alog: Drop largest and smallest $g = \lfloor \gamma n \rfloor$ and compute $\hat{\mu}$
   - **Influence function:** Large outliers have **no** effect on the estimator.
$$\psi(x) = \begin{cases} x, & -c < x < c \\ 0, & \text{otherwise} \end{cases}$$
2. **Winsorized Mean:** Modifies tail of distribution by replace
$$X_w = \frac{g \cdot X_{(g+1)} + X_{(g+1)} + \cdots + X_{(n-g)} + g \cdot X_{(n-g)}}{n}$$
3. **Summary**
   - No longer estimating population dist. mean $\int xf(x)dx$
   - Three quantities coincide only if population distribution is symmetric.
   - Trimmed/Winsorised mean is appropriate if we are interested in "typical" observation in the middle of the distribution

Measures of Scale

1. Median Absolute Deviation
$$P(|X - q_{f,0.5}| \le w) = 0.5 \text{ where } w := MAD(X)$$
   - Median of the distribution associated with $|X - q_{f,0.5}|$
   - MAD estimate $z_{0.75}\sigma$ if underlying distribution is Normal.
   - $\sigma \approx 1.4826 \times MAD(X)$
   - $MAD(X) = 0.6745\sigma$
2. **Interquartile Range:** $q_{f,0.75} - q_{f,0.25}$
$$\sigma \approx \frac{IQR(X)}{1.35}$$

Hypothesis Test

Skewness: method-of-moments estimator for the distribution skewness parameter
$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{[\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2]^{3/2}}$$
- close to 0 ⇒ low skewness (i.e. high symmetry).
- Positive values ⇒ **right**-skew
- Negative values ⇒ **left**-skew.

Kurtosis: thickness of the tails of a distribution.
$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^4}{[\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2]^2} - 3$$
- Positive ⇒ tails are "**fatter**" than those of a Normal.
- Negative ⇒ tails are "**thinner**" than those of a Normal.

Independent Samples Test
$$X_i \sim N(\mu_1, \sigma^2),\ i = 1, \ldots, n_1 \quad (1)$$
$$Y_j \sim N(\mu_2, \sigma^2),\ j = 1, \ldots, n_2 \quad (2)$$
$$T_1 = \frac{(\bar{X} - \bar{Y}) - 0}{s_p\sqrt{1/n_1 + 1/n_2}}$$

where
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$
- Under H_0, the test statistic $T_1 \sim t_{n_1+n_2-2}$.
$$CI : (\bar{X} - \bar{Y}) \pm t_{n_1+n_2-2,1-\alpha/2} \times s_p\sqrt{1/n_1 + 1/n_2}$$
- **Satterthwaite Approximation:** When we find variance not equal in two groups
$$T_{1,unpooled} = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$
The test statistic **still** follows a t distribution, but the **degrees of freedom are approximated**.

Paired Sample Test
$$T_2 = \frac{\bar{D} - 0}{s/\sqrt{n}}$$
where
$$s^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{(n-1)}$$
- Under $H-0$, the test statistic $T_2 \sim t_{n-1}$
- Confidence Interval
$$\bar{D} \pm t_{n-1,1-\alpha/2} \times s/\sqrt{n}$$

Wilcoxon Rank Sum: Begins by pooling the $n_1 + n_2$ data points and ranking them (Start from 1).
- Under H_0, the expected rank sum of group 1 is $E(R_1) = n_1 \times \frac{n_1+n_2+1}{2}$
- Used if both n_1 and n_2 are at least 10
- Observations should come from an underlying **continuous** distribution
- Test statistic W_1 follows a $N(0,1)$ distribution.

Wilcoxon Sign Test : Begin by ranking the $|D_i|$, then compute R_1, the sum of ranks for the positive D_i
- If rank sum R_1 is large, we expect that the pairs with $X_i > Y_i$ have a **larger difference** (in absolute values) than those with $X_i < Y_i$
- Under H_0, it can be shown that $E(R_1) = m(m+1)/4$
- If the number of non-zero D_i's is **at least 16**, then the test statistic W_2 follows a $N(0,1)$ distribution approximately.

MISC
- Applying multiple tests leads to a **higher Type I error**

ANOVA

Assumption of Normality
- If not hold ⇒ **Kruskal-Wallis test** (non-parametric version), to compare distributions between groups.

One-Way F-Test: Is there **any significant difference**, at 5% level, between the mean decomposition level of the groups?
$$Y_{ij} = \mu + \alpha_i + e_{ij} \sim N(\mu + \alpha_i,\ \sigma^2),\ i = 1, \ldots, k,\ j = 1, \ldots, n_i$$
- Setting $\sum_{i=1}^k \alpha_i = 0$, or
- Setting $\alpha_1 = 0$.
$$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{\overline{Y}}\right)^2}_{SST} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y_i}\right)^2}_{SSW} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} \left(\overline{Y_i} - \overline{\overline{Y}}\right)^2}_{SSB}$$
- Between Mean Square: $MS_B = \frac{SS_B}{k-1}$
- The Within Mean Square: $MS_W = \frac{SS_W}{n-k}$

Assumptions
- The **observations are independent** of each other.
- The errors are Normally distributed. (Check normality of $Y_{ij} - \overline{Y_i}$)
- The variance within each group is the **same**. (**rule-of-thumb**)

Comparing specific groups t-Test: if two particular groups i_1 and i_2 had different means

Contrast Estimation: comparison of a collection of l_1 groups with another collection of l_2 groups

Bonferroni: Perform m pairwise comparisons, to maintain the significance level of **each test** at α ⇒ we should perform each of the m tests/confidence intervals at α/m.

Kruskal-Wallis Procedure: Latter procedure is a generalisation of the Wilcoxon Rank-Sum test for 2 independent samples.
- Under H_0, the test statistic follows a χ^2 distribution with $k-1$ degrees of freedom.
- This test should only be used if $n_i \ge 5$ for all groups.

Linear Regression

- **constant variance assumption/homoscedascity:** $e_i \sim N(0, \sigma^2)$ implies
  - $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$, for $i = 1, \ldots, n$
  - $Var(Y_i|X_i) = Var(e_i) = \sigma^2$, for $i = 1, \ldots, n$
  - The Y_i are independent
  - The Y_i's are Normally distributed.
- ANOVA model:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SS_{Res}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SS_{Reg}}$$

- Estimate σ^2 with

$$\hat{\sigma^2} = \frac{SS_{Res}}{n-2}$$

- $R^2 = 1 - \frac{SS_{Res}}{SS_T} = \frac{SS_{Reg}}{SS_T}$ - The proportion of variation in Y_i, explained by the inclusion of X_i.
  - The larger the value of R^2 is, the better the model is.
  - Simply include more variables in the model will increase R^2, but it is undesirable
- Based on formulation $E(Y|X) = \beta_0 + \beta_1 X$, After estimating the parameters

$$\widehat{E(Y|X)} = \hat{\beta_0} + \hat{\beta_1} X \quad \text{(Fitted Line)}$$

  Thus we can **vary the values of** X to study how the mean of Y changes $\Rightarrow$ Dash Line for $100(1-\alpha)-$CI

Indicator Variables: If categorical variable has a levels, we will need $a-1$ columns of indicator variables

- The difference between the **models** is in the **intercept**. The other coefficients remain the same.

Interaction term: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3 + e$

- $X_3 = 1 \Rightarrow Y = (\beta_0 + \beta_3) + \beta_1 X_1 + (\beta_2 + \beta_4) X_2 + e$
- $X_3 = 0 \Rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$

Residual Diagnostics $r_i = Y_i - \hat{Y_i}$

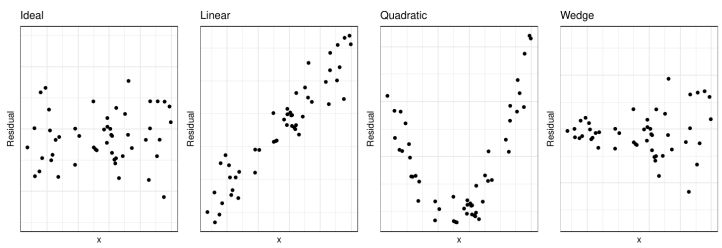
- contain only the information that our model **cannot explain**
- if the model is good $\Rightarrow$ contain random noise
- We can use residuals to
  - assess if the distributional assumptions hold
  - identify **influential points**
  - get direction on how to improve the model

Standardised Residuals: Use to check assumption of Normality

$$r_{i,std} = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

- Model fits well $\Rightarrow$ SR look similar to a $N(0,1)$
- Large SR $\Rightarrow$ potential outlier points.
- **Leverage** of a point (h_{ii}) measure of the potential influence of a point
- $h_{ii} \in (0, 1)$.
- Model with p parameters $\Rightarrow$ average $h_{ii} = p/n$.
- Points with $h_{ii} > 2 \times p/n \Rightarrow$ **high leverage** points.

Scatterplots:



1. Ideal. Residuals are **randomly distributed around zero**; there is **no pattern or trend** in the plot.
2. Probably appear if we were to plot residuals against a **new** variable (not currently in the model) $\Rightarrow$ should then include this variable in the model.
3. We should include a quadratic term in the model.
4. Indicates that we do not have homoscedascity $\Rightarrow$ either a transformation of the response, or weighted least squares.

Influential Point: The influence of a point on the inference can be judged by how much the inference changes **with and without** the point

Simulation

Strong Law of Large Numbers: If $X_1, X_2, \ldots, X_n$ are independent and identically distributed with $E(X) < \infty$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to E(X) \quad \text{with probability 1.}$$

- **No matter what distribution of** X **is**
- $\uparrow$ samples $\Rightarrow$ **sample mean** $\bar{X}$ converges to the desired value

Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be i.i.d., and suppose

- $-\infty < E(X_1) = \mu < \infty$
- $Var(X_1) = \sigma^2 < \infty$

$$\bar{X} \Rightarrow N(\mu, \frac{\sigma^2}{n}) \text{ or } \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \Rightarrow N(0,1)$$

- Use this theorem to obtain a CI for the expectation that we are estimating

Sample Estimates: Both the sample mean and sample standard deviation are **unbiased estimators**.

- $E(\bar{X}) = E(X)$
- $E(s^2) = \sigma^2$ where $s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$.
- $(1-\alpha)100\%$ CI for μ is $\bar{X} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$

When our goal is to estimate a probability p, we have to introduce a corresponding indicator variable X such that

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

In this case, the formula for the CI becomes

$$\bar{X} \pm z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

Monte-Carlo Integration

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) \mathrm{d}x$$

- $f(x)$ is a pdf, $X \sim f$
- It is critical that the support of the pdf is **the same as** the range of integration.

Type I Error: We falsely reject the null hypothesis 10% of the time if we perform it at 10% significance level

-*-*-*-*-*-*- PLEASE DELETE THIS PAGE! -*-*-*-*-*-*-

Information
Course:
Type: Cheat Sheet
Date: May 22, 2025
Author: QIU JINHANG
Link: https://github.com/jhqiu21/Notes

-*-*-*-*-*-*- PLEASE DELETE THIS PAGE! -*-*-*-*-*-*-