ST1131 Cheatsheet AY24/25 —— @Jin Hang

Exploratory Data Analysis

Variable: Any characteristic observed in a study.

- **Quantitative**: Observations take on numerical values.
  - Discrete: usually countable.
  - Continuous: infinitely many possible values
- **Categorical**: Each observation belongs to one of categories.
  - **Ordinal:** can be ordered, no specific quantitative values.
  - **Nominal:** have no specific ordering.
- **Difference:** Is distance between 2 points meaningful?
- Quant. $\to$ Cate.: divide into n ranges, count # in each range.

Frequency Table - Categorical

- **Proportion:** aka relative frequency. $\frac{\text{\# of obs. in 1 cat.}}{\text{Total \# of obs.}}$
- **Modal Category:** Category with highest frequency
- **Relative Frequencies:** Proportions and Percentages
- Summarizing: Modal category and its proportion

Bar Plots - Categorical

- Summarizing: Modal category and its proportion, Group of cat. with high/low proportions, Mention trends if ordinal

Histogram - Quantitative

- Skewed left/right: Left/right tail is longer
- Summarize: Unimodal/Bimodal/Multimodal, Skewed/ Symmetric, Outlier, Gap, Cluster
- (IQR, Range, $\bar{X} \pm 2s$) larger $\to$ variability$\uparrow$, less peaked

Describing Center

- **Mean:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Linear Transformation: $\bar{Y} = b\bar{X} + a$
  - Sensitive to outliers, unlike median
  - Symmetric and Bell-shaped: Report Mean
- **Median($X_{(0.5)}$):** $(\frac{n}{2})$th or mean.$(\frac{n}{2}, \frac{n}{2}+1)$
  - Robust to extreme observations
  - highly skewed, report median to summarize centre tendancy
- If $\bar{X} > X_{(0.5)}$, skew right. If $\bar{X} < X_{(0.5)}$, skew left.

Describing Variability

- **Range:** Max.-Min., Sensitive to outliers
- **Variance:** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Larger s means values are more spread out from mean.
- **Standard deviation:** $sd = \sqrt{S^2}$
  - Linear Transformation: $S_y^2 = b^2 s_x^2$, $S_y = |b| s_x$
  - **Empirical Rule:** $(\bar{X} \pm s) - 68\%$, $(\bar{X} \pm 2s) - 95\%$
- **Quartile:** (q_p) $100p\%$ of observations are below q_p
  - Lower quartile (Q_1), Median (Q_2), Upper quartile (Q_3)
  - Also known as **percentiles**. For continuous RV, the $100p$-th quartile, $q_p \Rightarrow P(X \le q_p) = p$
- **Inter-quartile Range (IQR):** $Q_3 - Q_1$, how spread out the "middle" of the sample is.
- Symm. $\to$ Mean, Variance. Skewed $\to$ Median, IQR.

Five-Number Summary: Min, Q_1, Q_2, Q_3, Max

Boxplot - Variability

- Identify Five-Number and Point out **outliers**
- **Outliers:** $< Q_1 - 1.5$ IQR or $> Q_3 + 1.5$IQR
- **Min/Max Whisker Reach:** Boundary of outliers
- **Upper/Lower Whisker:** Max/Min data within whisker reach
- Not portray certain features (i.e. mounds/gaps)
- If unimodal, can show skewness.
- **Summarizing:** Median, Outliers(# and which side of median), Compare medians, IQRs, Spread if > 1 boxplots, Skewness

Two Variables

- Response/Target Var.: variable on which comparisons are made
- Explanatory Var.: variable you believe the response depends on
- Sometimes unable to identify the role of variables $\Rightarrow$ Equally

Contingency Table - 2 categorical

- **Conditional Percentage** - % out of total
- Be careful of phrasing (Eg. Ppl w/o cancer of PMH users vs. PMH users of those w/o cancer)
- **Relative Risk** - Ratio of 2 percentages. (Eg. % of cancer in PMH users is 1.24 times the % of cancer in non-PMH users) Ratio is significantly different from 1 $\Rightarrow$ association between breast cancer and PMH usage.
- Raw difference is a statistic **help explore** the association

Scatter Plot - 2 Quantitative Variables

- Summarizing: Pos./Neg./No association, Linear/Trend, Range of y value, Constant variability, Outliers

Correlation: $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \in [-1, 1]$

- $r = \pm 1 \to$ Correlation is linear
- Correlation does not imply causation

Data Collection

- **Confounding Variable:** Occurs when two explanatory var. are associated with a response var., but are also associated with each other.
  - Observed variable that is included in the dataset.
  - difficult to tell which of the two explanatories is causing a change in the response.
- **Lurking Variable:** usually unobserved, that influences the association between the variables of primary interest
  - Typically not measured in the study
- **Experimental Study:** Assign subjects (experimental units) to certain experimental conditions(treatments) and observe resp.
  - Control lurking var. by randomly assigning the treatment.
  - More confident to determine causality between exp./resp.
- **Observational Study:** Explanatory and response variable observed for subjects. No treatments.

Sample Survey

1. Identify the Population
2. Compile **Sampling Frame:** Where sample is from
   - Ideally, sampling frame lists all subjects in population.
3. **Sampling design:** How to choose subjects from frame
4. Collect data from the chosen sample.

Simple Random Sample: Same chance of being chosen

- Ensures it is representative of the general population.
- Allow us to make inferences about the population.
- Cluster Random Sampling, Stratified Random Sampling...

Sources of Bias in Sample Survey:

- **Sampling Bias:** Sample not random or undercoverage
- **Non-response Bias:** No response from subject
- **Response Bias:** Incorrect resp./misleading qns

A large sample size does not guarantee an unbiased sample.

Convenience Sample: selected based on ease of access. E.g. outside a mall or at an MRT station.

Volunteer samples: People are encouraged to participate in the survey via a flyer or email. This can yield **incorrect inferences**.

Elements of Good Experimental Study:

- Control comparison group
- Randomization: Eliminate lurking variables
- Blinding the study: Placebo

Sample Space: Set of all possible outcomes of a random phem.

Event: Subset of the sample space S.

- corresponds to a particular or a group of possible outcomes.
- Complement(A^C) consists of all outcomes $\notin$ A.
- **Mutually Exclusive:** $A \cap B = \emptyset$
- **Independent:** $A \perp B \Rightarrow P(AB) = P(A) \cdot P(B)$
  1. **独立和互斥的关系:** 独立不互斥, 互斥不独立
  2. S and $\emptyset$ are independent of any other event.
  3. If $A \perp B$, then $A \perp B'$, $A' \perp B$, and $A' \perp B'$.
  4. $A \perp B \Rightarrow P(A) = P(A|B)$ & $P(B) = P(B|A)$
  5. $P(A) = 0/1 \Rightarrow A$ is **independent** of any B
  6. $0 < P(A), P(B) < 1$, 若AB互斥/包含 $\Rightarrow$ 不独立
- **Intersection:** $A \cap B$ contains elements common to both

1. $A \cap A' = \emptyset$	6. $(A \cup B)' = A' \cap B'$
2. $A \cap \emptyset = \emptyset$	7. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
3. $A \cup A' = S$	8. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
4. $(A')' = A$	9. $A \cup B = A \cup (B \cap A')$
5. $(A \cap B)' = A' \cup B'$	10. $A = (A \cap B) \cup (A \cap B')$

$A \subset B$ if all elements in event A are in event B, if $A \subset B$ and $B \subset A$ then $A = B$. We assume **contained** means proper subset.

Probability Axioms

- $P(A \cup B) = P(A) + P(B) \le 1$
- $P(AB) \le P(A)$ or $P(B)$
- $P(A) = P(A \cap B) + P(A \cap B')$
- $A \subset B \Rightarrow P(A) \le P(B)$
- If $P(A) \ne 0, A \perp B$ if and only if $P(B \mid A) = P(B)$

Properties of Independent Events

1. $\mathrm{P}(B|A) = \mathrm{P}(B)$ and $\mathrm{P}(A|B) = \mathrm{P}(A)$
2. A and B cannot be mutually exclusive if they are independent, supposing $\mathrm{P}(A), \mathrm{P}(B) > 0$
3. A and B cannot be independent if they are mutually exclusive
4. Sample space S and empty set $\emptyset$ are independent of any event
5. If $A \subset B$, then A and B are dependent unless $B = S$.

Probability

- **Conditional:** $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$, if $\Pr(A) \ne 0$
- **Multiplicative:** $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A) = \mathrm{P}(B)\mathrm{P}(A|B)$
- **LoTP:** $\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(B \cap A_i) = \sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B|A_i)$
- **Bayes:** $\mathrm{P}(A_k|B) = \frac{\mathrm{P}(A_k)\mathrm{P}(B|A_k)}{\sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B|A_i)}$

Epidemiological Terms

- **Sensitivity:** $P(+|D)$, Given has disease, prob. of positive test.
- **Specificty:** $P(-|\bar{D})$ Given has no disease, prob. of neg. test
- **Prevalence:** $P(D) = \frac{\text{\# of people with disease}}{\text{Total population}}$
- **Positive Predictive Value:** $P(D|+)$

Random Variable

Discrete

- $\sum_i P_i = 1$
- $\mu = \sum_x x P_x$
- $\sigma^2 = \sum_x P_x (x - \mu)^2$
- Visualize: Bar Plot - Width of each rectangle is identical, but the height is proportional to p_x

Continuous Random Variables: Distribution represented by **probability density function**, area under curve = 1.

- $\mu = \int x f(x)\,\mathrm{d}\,x$
- $\sigma^2 = \int (x - \mu)^2 f(x)\,\mathrm{d}x$

Mean

- If $x_1, ..., x_n$ have same prob. distri., mean of these variables ($\bar{X}$) is a random variable where $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu$
- a large number of observations from a population, sample mean of those observations would be close to the mean of that probability distribution.

Variance

- $Var(\bar{X}) = \frac{\sigma^2}{n}$

Binomial Distribution $\mathrm{Bin}(n, p)$

1. n independent trials with 2 outcomes
2. Each trial has probability of p to succeed

Binomial Random Variable - # of successes in n trials

- $\mathrm{Bernoulli}(p) \Leftrightarrow \mathrm{Bin}(1, p) \Rightarrow \sum \mathrm{Bernoulli}(p) = \mathrm{Bin}(n, p)$
- $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
- $E(X) = np$
- $Var(X) = np(1-p)$

Poisson Distribution: k is num(occurrences) of events

- $P(X=k)=\frac{e^{-\mu}\mu^k}{k!}$
  - λ is the expected no. of events per time unit
  - $\mu=\lambda t$ is the expected no. of events over time period t.
- $E(X)=\lambda$ • $V(X)=\lambda$ • $X\sim \text{Poisson}(\lambda)$
- $n\to\infty, p\to 0$, $\text{Bin}(n,p)\to \text{Poisson}(np)$ if $n\geq 20$ and $p\leq 0.05$, or if $n\geq 100$ and $np\leq 10$.

Normal(Gaussian) distribution: $X\sim N(\mu,\sigma^2)$

- If $d>0, P(X\leq \mu-d)=P(X\geq \mu+d)$.
- $q_{1-p}=2\mu-q_p$
- $N(0,1)$ between -1 and 1 $\sim 68\%$
- $N(3,4)$ between -1 and 7 $\sim 95\%$
- $X,Y\sim N(\mu,\sigma^2)\Rightarrow aX+bY\sim N(a\mu_X+b\mu_Y, a^2\sigma_X^2+b^2\sigma_Y^2)$
- $X\sim N(\mu,\sigma^2)\Rightarrow Z=\frac{X-\mu}{\sigma}\sim N(0,1)$ [Z**-Score of** X]
- $Z\sim N(0,1)\Rightarrow -Z\sim N(0,1)\Rightarrow \sigma Z+\mu\sim N(\mu,\sigma^2)$
- **Sensitivity** p: $\exists$ cut-off value C s.t. $P(X\geq C)=p$
- **Approximation:** For n is moderately large and p is not close to 0 or 1, if $np(1-p)\geq 5$, $X\sim \text{Bin}(n,p)\sim N(np, np(1-p))$

Sampling Distribution

- **Data Distribution** - Distribution of some observations from a single sample. Larger n, closer data distribution to the pop. distribution.
- **Sampling Distribution** - Distribution of $\bar{X}$ and $\hat{p}$
- **Central Limit Theorem** - Suppose there are independent observations that form a distribution (not necessarily normal) with mean μ and variance σ^2 and sample size n is large, then sample mean $\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$

Sample Proportion $\hat{p}=\frac{X_1+\ldots+X_n}{n}$

- **Population Proportion** - p that we want to estimate
- **Population Distribution** - $Ber(p)$ where $\mu=p$ and $\sigma^2=p(1-p)$
- When $np(1-p)\geq 5$, $\hat{p}\sim N(p,\frac{p(1-p)}{n})$ approximately by CLT

Sample Mean $\bar{X}$**:** when population distribution is

1. normal
   - $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}$
   - $\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$ exactly
2. not normal
   - $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}$
   - When $n\geq 30$, $\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$ approximately by CLT

Statistics Inference: There are two main types of inference

- estimation of population parameters, and
- testing hypotheses.

Point Estimate A single number that is best guess for pop. para.

- $\bar{X}\to\mu$, $s^2\to\sigma^2$, $\hat{p}\to p$, $X_{0.5}\to q_{0.5}$
- Does not show how close they are to true value

Interval Estimate Interval of numbers within which the parameter value is believed to fall.

- Indicates precision by an interval of nums around point est.
- The interval is made up of numbers that are the most believable values for the unknown parameter, based on the data observed.

Confidence Intervals: CI = Point estimate $\pm$ Margin of error

- **Margin of error** measures how accurate the point estimate is likely to be in estimating a parameter.
- **Standard Error** - (SE) Estimated sd of **sampling distribution**
- We have X% confidence that p falls in the interval ..
- We could have 100%-CI without any sample $\Rightarrow$ do use that!
- Sample Size $\uparrow\Rightarrow$ SE $\downarrow\Rightarrow$ Length(CI) $=2\times q_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\downarrow$

Find CI given confidence lvl. (x)**:**

1. Find $\hat{p}$ and check $n\hat{p}(1-\hat{p})\geq 5$
2. Let $\alpha=1-x$
3. $\text{CI}=\hat{p}\pm q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=\hat{p}\pm q_{1-\frac{\alpha}{2}}\text{SE}$

Determine sample size (n) **before study:**

1. Decide confidence level (x) and width of CI (D)
2. $n\geq(\frac{2q_{1-\frac{\alpha}{2}}}{D})^2p(1-p)$ where $p=\frac{1}{2}$

Confidence Interval for Mean

- σ^2 of pop. will affect Len of interval, but we cannot change it.

t-distribution For $\bar{X}\sim N(\mu,\sigma^2/n)$, $\frac{\bar{X}-\mu}{s/\sqrt{n}}\sim t_{n-1}$

- Has thicker tails and more variability than that of $N(0,1)$
- (t_{df}) Approaches $N(0,1)$ as $df\uparrow\Rightarrow df\geq 30\Rightarrow t_{df}\sim N(0,1)$

Find CI given confidence lvl. (x)**:**

1. **Assumptions:** Sample is random (**not robust**, crucial); Data distribution symmetric or n is big(**robust**)
2. $\text{CI}=\bar{X}\pm t_{n-1;1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$. **Note:** If n is large enough and σ is known, then the margin of error is $z_{\alpha/2}(\sigma/\sqrt{n})$ instead.

Robustness of Assumptions: A statistical method is said to be **robust** with respect to a assumption if it performs adequately even when that assumption is modestly violated.

Determine sample size (n) **before study:**

1. Decide confidence level (x) and width of CI (D)
2. $\text{Length(CI)}=2\times t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\leq D\Rightarrow n\geq(\frac{2t_{n-1;1-\alpha/2}\cdot s}{D})^2$
3. We don't know n, $s\Rightarrow$ use $n\geq(\frac{2q_{1-\alpha/2}\cdot s}{D})^2$ For s, look for similar studies. Ensure $n\geq 30$.

Hypothesis Testing

- **Test statistic:** How far point estimate falls from guess
- **Null distribution:** Distribution of test stat. under H_0
- **p-value:** How unlikely observed value is, if H_0 is true
- **Significance level**(α)**:** Tells us how strong the evidence should be. Reject H_0 if p-value $\leq\alpha$
  Test is statistically significant when we reject H_0

Two Errors: Increase sample size to reduce both errors

- Type I: Reject H_0, but H_0 is true
  $\alpha=P_{H_0}(z\geq c_{\text{reject}})\Rightarrow$ The smaller α the better!
- Type II: Do not reject H_0, but H_0 is false
  $\beta=\alpha=P_{H_1}(z\geq c_{\text{reject}})\Rightarrow$ The smaller β the better!
- **Power of Test** $=1-\beta$, probability of correctly rejecting H_0, when it is in fact false.
- Cannot reduce both types of errors simultaneously.
- $\alpha\downarrow\Rightarrow\beta\uparrow$, but impossible $\alpha=0,\beta=1$

One sample, Proportion

1. Assumptions: Categorical, Random, $np_0(1-p_0)\geq 5$
2. Hypothesis: $H_0: p=p_0$ and $H_1: p(>/</\neq)p_0$
3. Test statistic: $z=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ and $z\sim N(0,1)$
4. **p-value:** Right sided - $P(Z\geq z)$; 2-sided - $2P(Z\geq z)$
5. Reject H_0 if p-value $\leq\alpha$.

One sample, Mean

1. Quantitative, Random, Approx. normal (or $n\geq 30$)
2. Test statistic: $T=\frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$ and $T\sim t_{n-1}(0,1)$
3. Result of 2-sided test for mean is same as using CI

Two sample, Independent, Equal variance

1. Assumptions: Quantitative, Random, Independent samples, Pop. distri. is approx. normal (or n is large enough), Equal variance test >0.05
2. Hypothesis: $H_0:\mu_1=\mu_2$ and $H_1:\mu_1><\neq\mu_2$
3. Test statistic: $T=\frac{(\bar{X}-\bar{Y})-0}{\text{SE}}\sim t_{n_1+n_2-2}$ where $\text{SE}=s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ and $S_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ (Pooled estimate of common variance)

Two sample, Independent, Unequal variance

1. Assumptions: Same, except pop. var. is different
2. Test statistic: $T=\frac{(\bar{X}-\bar{Y})-0}{\text{SE}}\sim t_{df}$ where $\text{SE}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ and df needs R $\Rightarrow$ Use R to get p-value and df

Two sample, Dependent

- 2 samples are dependent $\leftrightarrow$ Each obs. has matching pair (Eg. Before and after)
- Take difference of matched observations and compare mean of difference with 0. Similar to 1 sample test.

Normality Assumption: Sample distribution is approximately Normal $\Rightarrow$ high probability that the population follows it as well.

QQ Plot

- Right $\swarrow$, Left $\nearrow\Rightarrow$ Longer; Right $\nwarrow$, Left $\searrow\Rightarrow$ Shorter
- Summary: longer/shorter tail than normal, occurs on which side of the mean

Shapiro-Wilk Test: Good for small samples only.

- H_0 : Sample is from a Normal distribution. H_1 : Not normal.

Regression Model

- **Linear** refers to the linearity in the parameters.
  - $Y=\beta_0+\beta_1\sin(X_1)+\beta_2\log(X_2)+\beta_3e^{X_3}+\varepsilon$ is linear
  - $Y=\beta_0\sin(\beta_1X)+\varepsilon$, $Y=\beta_0e^{\beta_1X}+\varepsilon$ are non-linear

Simple Linear Regression: $Y=\beta_0+\beta_1x+\epsilon$

- Response Variable: Y, Explanatory Variable: x, Regressor: X
- Assumptions: Random data, Relationship is linear, $\epsilon\sim N(0,\sigma^2)$ $\Rightarrow Y\sim N(\beta_0+\beta_1x,\sigma^2)$ (Check if resp. is symmetric)
- We check the assumptions **after** fitting the model
- $\hat{\sigma}$ = Residual standard error in model summary
- **Ordinary Least Square Estimate** - Line with least sum of square residuals. $ss_{Res}=\sum_{i=1}^n e_i^2$ where $e_i=y_i-\hat{y}_i$
- **Interpretation:** Point estimates of Y mean at different X val.
- **Interpolation:** Estimate not observed. Within range
- **Extrapolation:** Estimate that's outside range. Avoid!

T-test and F-test

- **T-test:** Test significance of 1 coefficient
- **F-test:** Test significance of entire model
  $\Rightarrow$ F-test not significant $\Rightarrow$ new model without any regressor $Y=\beta_0+\varepsilon$ reduce to **intercept model** $\hat{Y}=\hat{\beta}_0$
- **Simple:** only one F-test $\Leftrightarrow$ T-test. **Multi:** >1 T-test

1. **Assumptions:** Same as building model
2. **T-test:** $H_0:\beta_1=0$(Regressor X is not signif.), $H_1:\beta_1\neq 0$
3. **F-test:** H_0 : All coeff.$\backslash\beta_0=0$, H_1 : At least 1 coeff.$\neq 0$.
4. $t=\hat{\beta}_1/SE(\hat{\beta}_1)$: Check from summary. Null distri: $t_{\text{n - \# of coeff}}$
5. **F-stat:** Simple: $F=t^2$. Null Distri: $F_{\text{\# of coeff; n - \# of coeff}}$

Regression Diagnostics

- Scatter plot Y againest X
  - Linearity assumption violated: add higher order terms in X.
  - Variance not constant: Try $\ln(Y)$, $\sqrt{Y}$ or $1/Y$.
  - Transformation will change interpretation the coefficient β_1.
- Residual (raw residual $e_i=Y_i-\hat{Y}_i$) plots to check for
  - normality assumption $\Rightarrow$ Histogram and QQ plot
  - non-constant variance and the need to transform Y.
  - need to add higher order terms in X.
  $\Rightarrow r_i(SR)$ on y-axis vs. $\hat{Y}_i/X$ on x-axis with interval $(-3,3)$
  $\star$ Funnel Shape $\to$ non-constant variance; Curve $\to$ Linearity
- Standard Residual (SR) $=\frac{Y-\hat{Y}}{\text{standard error of }(Y-\hat{Y})}$

- **Outliers:** Identified by the residuals, far from rest data point.
- **Influential Point:** affects the parameter estimates greatly.
  - Outlier may or may not be influential
  - Points with a large Cook's distance (measures the effect of deleting a given observation).

Coefficient of Determination(R^2): Goodness of fit

- **Interpretation:** proportion of total variation of the response (about the sample mean Y) that is explained by the model.
- $|\text{Cor}(x,y)| = \sqrt{R^2} = R$, $\hat{\beta_1} < 0 \to \text{Cor(x, y)} < 0$
- More variables $\to R^2 \uparrow \to$ use Adjusted $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$
- R^2 cannot compare two model, use R^2_{adj}

Multivariate Linear Regression Regression Function with Categorical Var:

- **Indicator Variable:** 1 if cat. is observed. 0 otherwise.
- **Reference Category:** The category not in equation

Eg. $Y = \beta_0 + \beta_1 x_1 + \beta_2 I(x_2 = Auto) + \epsilon$

- Auto: $Y = \beta_0 + \beta_1 x_1 + \beta 2 + \epsilon$
- Manual: $Y = \beta_0 + \beta_1 x_1 + \epsilon$

Interaction between variables:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 I(x_2 = Auto) + \beta_3 x_1 I(x_2 = Auto) + \epsilon$

Interpretation of Coefficients:

$$y_1 = \beta_0 + \beta_1 X_1 + \beta_2 I(X_2 = 1)$$

- For any fixed X_2, when X_1 increases by 1 unit, the y will increase by β_1 unit
- For any fixed X_1, the y of type 1 ($X_2 = 1$) is β_2 more than the one of type 0 ($X_2 = 0$)

$$y_2 = \beta_0 + \beta_1 X_1 + \beta_2 I(X_2 = 1) + \beta_3 X_1 \times I(X_2 = 1)$$

- For type 1 ($X_2 = 1$) when X_1 increases by 1 unit, y will increase by $(\beta_1 + \beta_3)$ unit. For type 0 ($X_2 = 0$) when X_1 increases by 1 unit, y will increase by $\beta_1 unit$.

Reference

Sample Statistics: Sample resembles Population!

Population parameters: Values Computed (μ, σ, p)

Statistic: Suppose a random sample of n, $(X_1, \ldots, X_n)$ has been taken. A function of $(X_1, \ldots, X_n)$ is called a **statistic**.

1. **Sample mean:** $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ($\bar{X}, S$ are **Random Var.**)
2. **Sample variance:** $S^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$

Mean and Variance of $\bar{X}$:

- $\mu_{\bar{X}} = E(\bar{X}) = \mu_X$ • $\sigma^2_{\bar{X}} = V(\bar{X}) = \frac{\sigma^2_X}{n}$
- $E\left(S^2\right) = E\left[\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2\right] = \sigma^2$
- $\bar{X}$ estimates μ_X, $n \uparrow$. $\sigma^2_X/n \downarrow \Rightarrow \mu_X \to \bar{X}$.

CLT: $n \to \infty \Rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to Z \sim N(0,1) \Leftrightarrow \bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right)$

1. $\lim_{n\to\infty} P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x)$
2. $X_1, X_2, \ldots, X_n$ independent and $N\left(\mu, \sigma^2\right)$, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ or $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ regardless n.

t-distribution: $Z \sim N(0,1)$, $U \sim \chi^2(n)$. If Z and U independent, then $T = \frac{Z}{\sqrt{U/n}} \sim t(n)$

Properties

1. The t-distribution approaches $N(0,1)$ as $n \to \infty$ i.e. $n \geq 30$
2. If $T \sim t(n)$, then $E(T) = 0$ and $V(T) = n/(n-2)$ for $n > 2$.
3. $X_i \sim N(\mu, \sigma^2)$, $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

If $X_1, \ldots, X_n$ independent & identically distributed $X_i \sim N(0,1)$, $\bar{X}$ is the sample mean and S^2 is the sample variance, then

- $n\bar{X} \sim N(0,n)$. • $\frac{\sqrt{n}\bar{X}}{S} \sim t_{n-1}$.

Estimation

Unbiased Estimator Let $\hat{\Theta}$ be an **unbiased** estimator of θ. Then $\hat{\Theta}$ is a random variable based on the sample s.t. $E(\hat{\Theta}) = \theta$

- $\bar{X}$ is a good estimator of μ
- $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^n \mu_X = \mu_X$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$, $E\left(S^2\right) = \sigma^2$
- $\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \text{var}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2_X = \frac{\sigma^2_X}{n}$

z_α: Number with upper-tail prob. of α s.t. $P\left(Z > z_\alpha\right) = \alpha$.

Maximum Error of Estimate: $\bar{X} \neq \mu \Rightarrow \bar{X} - \mu$ measures difference between estimator and the true value of the parameter. If population is normal or n is large, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows a standard normal or an approximately standard normal distribution.

$P\left(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(|\bar{X} - \mu| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

$\Rightarrow E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Determine Sample Size: $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq E_0 \Rightarrow n \geq \left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2$

Interval Estimator: For (a,b) you are fairly certain the parameter of interest lies in, quantified by confidence level $(1-\alpha)$ s.t. $P(a < \mu < b) = 1 - \alpha$. (a,b) is $(1-\alpha)$ confidence interval.

- $(1-\alpha)$ confidence interval can be written as $\bar{X} \pm E$.
- $\bar{X} \pm E$ has probability $(1-\alpha)$ of containing μ
- Once computed, μ is either in it or not $\Rightarrow$no more randomness.
- n is large when $n \geq 30$

Comparing Two Population: Confidence Intervals for $\mu_1 - \mu_2$

- **Independent samples**: complete randomization.
- **Matched pairs samples**: randomization between pairs.

Pooled estimator(S_p^2): σ^2 can be estimated by the **pooled sample variance** $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ with S_1^2 and S_2^2 being the sample variances of the first and second samples respectively.

Roughly assume equal variance if $1/2 \leq S_1/S_2 \leq 2$ not sensitive to small difference between population var.

Paired Data: For $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)$

- X_i and Y_i are dependent. • $\left(X_i, Y_i\right)$ are independent
- Define random sample $D_i = X_i - Y_i, \mu_D = \mu_1 - \mu_2$.
- **Small** and Normal: $\bar{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$; **Large:** $\bar{d} \pm z_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$

Hypothesis Tests

- **Null**: Try to it's false (Type I may happen reject null.), makes an assertion that a parameter equals to some constant.
- **Ater.**: Prove to be true, againest **null.** Reject $H_0 \Rightarrow$ Concl. H_1 Type II occur if do not reject null
- Reject null. $\Rightarrow$ enough evidence to support alternative.

Type I/II Error	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	Type I error
H_0 is false	Type II error	Correct Decision

- The Type I error: serious $\to$ control P(Type I)
- Thus prior to conducting a hypothesis test, we set the significance level α to be small, typically at $\alpha = 0.05$ or 0.01
- Did not "prove" that H_0 is true $\Rightarrow$ Not accept

p-value(observed level of significance): Probability of obtaining a test statistic at least as extreme ($\leq$ or $\geq$) than the observed sample value, given H_0 is true.

$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
$P(\lvert Z\rvert > \lvert z\rvert)$	$P(Z < -\lvert z\rvert)$	$P(Z > \lvert z\rvert)$

- p-value $< \alpha$, reject H_0; p-value $\geq \alpha$, do not reject H_0

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \lvert z\rvert)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\lvert z\rvert)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z\rvert)$

Level of significance:

$\alpha = P($ Type I error $) = P($ Reject $H_0 \mid H_0$ is true $)$

Power of Test: $1 - \beta = P($ Reject $H_0 \mid H_0$ is false $)$, where $\beta = P($ Type II error $) = P($ Do not reject $H_0 \mid H_0$ is false $)$

$\alpha \uparrow \beta \downarrow \quad \alpha + \beta \neq 1$

Rejection Region:

- $H_1 : \mu \neq \mu_0$, $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
- $H_1 : \mu < \mu_0$, $z < -z_\alpha$
- $H_1 : \mu > \mu_0$, $z > z_\alpha$

Step 1: Set the null and alternative.
Step 2: Set $\alpha = 0.05$
Step 3: Use test with test statistics, determine rejection region
Step 4: Calculate observed value use Step 3 distribution
Step 5: Reject/Not Reject

R syntax

Dataframe: Similar to the Matrix object but columns can have different modes.

```
v <- c(1:6) // 1 2 3 4 5 6
m <- matrix(v, nrow=2, ncol=3)          // 1 3 5
m <- matrix(v, nrow=2, ncol=3, byrow=T) // 1 2 3
ab_row <- rbind(c(1,2,3),c(4,5,6))      // 1 2 3
ab_col <- cbind(ab_row, c(9, 10))       // 1 2 3 9
rt = data.frame(response, treatment)
read.csv(...) / read.table(...)
scan(..., what = "character") // Only Read one type

prop.table(table(lung$Gender)) // frequency table (prob)
lung$Gender <- ifelse(lung$Gender=="0", "Female", "Male")
barplot(table(lung$Gender), ylab="..", xlab="..", col=c
     (2,5), main="header..") // bar plot 'col' is color
pie(table(lung$gender), col=c(2,5), main="....")
hist(mark, prob=TRUE, col=2, xlab=".", ylab=".", main=".")
boxplot(mark, ylab = ".", main = ".", col=5)
abline(h = median(mark), col = "red") // add a line

pnorm(1800, mean = 1500, sd = sqrt(90000))
pnorm(1630, mean = 1500, sd = sqrt(90000), lower.tail =
FALSE)
```

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

Information
Course: ST1131 Introduction to Statistics and Statistical Computing
Type: Cheat Sheet
Date: May 22, 2025
Author: QIU JINHANG
Link: https://github.com/jhqiu21/Notes

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