ST1131
AY21/22 Sem 2
github.com/jasonqiu212

01. Exploratory Data Analysis

- **Quantitative Variable**: Discrete vs. Continuous
- **Categorical Variable**: Ordinal vs. Nominal
- Difference: Is distance between 2 points meaningful?

Single Variable

Frequency Table - Categorical

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

Bar Plots - Categorical

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

Histogram - Quantitative

- Skewed left/right: Left/right tail is longer
- Summarizing: Unimodal/Bimodal/Multimodal, Skewness, Outlier

Describing Center

- **Mean** - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Linear Transformation: $\hat{Y} = b\hat{X} + a$
  - Sensitive to outliers, unlike median
- **Median** - $X_{(0.5)}$
- If $\bar{X} > X_{(0.5)}$, skew right. If $\bar{X} < X_{(0.5)}$, skew left.

Describing Variability

- **Range** - Sensitive to outliers
- **Variance** - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Standard deviation** - $sd = \sqrt{S^2}$
  - Linear Transformation: $S_y^2 = b^2 s_x^2$ $S_y = |b| s_x$
- **Inter-quartile Range (IQR)** - $Q_3 - Q_1$
- **Quantile** - (q_p) $100p\%$ of observations are below q_p
- Lower quartile (Q_1), Median (Q_2), Upper quartile (Q_3)
- Symm. $\rightarrow$ Mean, Variance. Skewed $\rightarrow$ Median, IQR.

Boxplot - Variability

- Includes: Min, Q_1, Q_2, Q_3, Max
- **Outliers** - $< Q_1 - 1.5IQR$ or $> Q_3 + 1.5IQR$
- **Max/min Whisker Reach** - Boundary of outliers
- **Upper/lower Whisker** - Min/max obs. exc. outliers
- If unimodal, can show skewness.
- Summarizing: Median, Outliers, Compare medians and IQRs if > 1 boxplots

Two Variables

- Response Variable vs. Explanatory Variable

Bar Plots - 2 categorical

Contingency Table - 2 categorical

- **Conditional Percentage** - % out of total
- **Join Percentage** - % out of some group. Use explanatory as group.
- 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)

2 Boxplots - 1 Categorical and 1 Quantitative

Scatter Plot - 2 Quantitative Variables

- Summarizing: Pos./neg. association, Linear, Constant variability, Outliers

Correlation - $r \in [-1, 1]$

- $r = \pm 1 \rightarrow$ Correlation is linear

02. Data Collection

- **Confounding Variable** - Related to exp. and resp. variable. Confounds their association. Observed.
- **Lurking Variable** - Unobserved
- **Experimental Study** - Assign subjects (or **experimental units**) to **treatments** and observe response variable
- **Observational Study** - Explanatory and response variable observed for subjects. No treatments.

Sample Survey

1. Compile **sampling frame** - Where sample is from
2. **Sampling design** - How to choose subjects from sampling frame
   - **Simple Random Sample** - Each sample has same chance of being chosen

Sources of Bias in Sample Survey:

- **Sampling Bias** - Sample not random or undercoverage
- **Non-response Bias** - No response from subject
- **Response Bias** - Subject does not answer truthfully

Elements of Good Experimental Study:

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

03. Probabililty

- **Sample space** - (S) Set of all possible outcomes
- **Event** - (E) Subset of sample space
- $P(A) = \frac{\#\text{ of outcomes in A}}{\text{Total }\#\text{ of possible outcomes}}$

Axioms of Probability

1. $0 \le P(A) \le 1$
2. $P(S) = 1$
3. If A and B are mutually exclusive (or **disjoint**), then $P(A \cup B) = P(A) + P(B)$ and $P(A \cap B) = 0$
4. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

For any events A and B:

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) = P(A \cap B) + P(A \cap B^c)$
- A and B are **independent** if $P(A \cap B) = P(A)P(B)$

Conditional Probability

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Law of Total Probabililty

$P(A) = P(A \cap B_1) + ... + P(A \cap B_n)$

Bayes' Theorem

$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + ... + P(A|B_n)P(B_n)}$

Epidemiological Terms:

- **Sensitivity** - Given has disease, prob. of positive test
- **Specificty** - Given has no disease, prob. of neg. test
- **Prevalence** - $\frac{\#\text{ of people with disease}}{\text{Total population}}$

04. Random Variables

- **Random variable** - Unpredictable outcome of exp.
- **Probability distribution of random variable** - Possible values and their probabilities

Discrete Random Variables

- **Probability distribution**: (P_x) Prob. for each possible x. Sum of all $P_x = 1$

Mean

- aka **expected value**. $\mu = \sum_x xP_x$
- Mean of observations approach μ with lots of obs.
- Linear transformation: $E(Y) = a\mu + b$ and $E(a_1x_1 + ... + a_nx_n) = a_1\mu_1 + ... + a_n\mu_n$
- If $x_1, ..., x_n$ have same prob. distri., mean of these variables $(\bar{X})$ is a random variable where $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu$

Variance

- $\sigma^2 = \sum_x P_x(x - \mu)^2$ and sd: σ
- Linear transformation: $Var(Y) = b^2Var(X) = b^2\sigma^2$ and $Var(a_1x_1 + ... + a_nx_n) = a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2$
- Likewise, $Var(\bar{X}) = \frac{\sigma^2}{n}$

Continuous Random Variables

- **Probability distribution**: Represented by **probability density function**. Area under curve = 1.
- Mean and variance have same properties as discrete

Binomial Distribution

- **Combinations** - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

3 Conditions for binomial distribution:

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

Binomial Random Variable - # of successes in n trials

- Follows distribution: $Bin(n, p)$
- **Bernoulli Distri.** - $Bin(1, p)$. Sum of Ber. $\sim$ Bin.

Binomial Formula: Suppose $X \sim Bin(n, p)$

- $P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$
- $E(X) = np$
- $Var(X) = np(1-p)$

Normal Distribution

- Symmetric about μ, bell-shaped, $X \sim N(\mu, \sigma^2)$
- **Standard Normal Distribution** - $N(0, 1)$
- Lin. transf. of Normal Random Variable: Same behavior for mean and variance
- Approximate binomial distribution using $N(np, np(1-p))$ when $np(1-p) \ge 5$

Z-score - $z = \frac{x - \mu}{\sigma} \sim N(0, 1)$

- Outlier: Any observation with z-score of > 3 or < -3

QQ Plot

- Purpose: To see if data follows $N(\mu, \sigma^2)$
- Compare right/left tails with normal

05. Sampling Distribution

- **Data Distribution** - Distribution of some observations
- **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 $\hat{p}$

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

Sample Mean $\bar{X}$

when population distribution is normal:

- $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$
- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ exactly

when population distribution is not normal:

- $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$
- When $n \ge 30$, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ approximately by CLT

06. Confidence Intervals

- In long run, 95% of intervals will contain population proportion/mean.

Point Estimate

- $\bar{X} \to \mu$ and $\hat{p} \to p$
- Does not show how close they are to true value

Confidence Interval for Proportion

- CI = Point estimate $\pm$ Margin of error
- **Standard Error** - (se) Estimated sd of sampling distri.

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. CI $= \hat{p} \pm q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Determine sample size (n) before study:

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

Confidence Interval for Mean

- **t-distribution** - (t_{df}) Approaches $N(0,1)$ as $df \uparrow$

Find CI given confidence lvl. (x):

1. Assumptions: Sample is random (**not robust** - crucial), Data distribution symmetric (or n is big)
2. CI $= \bar{X} \pm t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

Determine sample size (n) before study:

1. Decide confidence level (x) and width of CI (D)
2. $n \geq (\frac{2sq_{1-\frac{\alpha}{2}}}{D})^2$
3. For s, look for similar studies. Ensure $n \geq 30$.

07. Hypothesis Testing

- **Null hypothesis** (H_0) vs. **Alternative hypothesis** (H_1)
- **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) Reject H_0 if p-Val $\leq \alpha$
- Test is **statistically significant** when we reject H_0
- **Type I Error** - Reject H_0, but H_0 is true
- **Type II Error** - Do not reject H_0, but H_0 is false
- Increase sample size to reduce both errors

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: If right sided, $P(z \geq$ Test stat.$|z \sim N(0,1))$. If 2-sided, $2 * P(z \geq$ Test stat.$|z \sim N(0,1))$
5. Conclusion: Reject H_0 if p-Val $\leq \alpha$. Else, cannot reject

One sample, Mean

1. Assumptions: Quantitative, Random, Data distri. is approx. normal (or $n \geq 30$)
2. Hypothesis: $H_0 : \mu = \mu_0$ and $H_1 : \mu >< \neq \mu_0$
3. Test statistic: $T = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$ and $T \sim t_{n-1}(0,1)$
4. p-Value and Conclusion: Same as proportion

- 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: $T = \frac{\bar{X}-\bar{Y}}{se}$ where $se = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 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**) and $T \sim t_{n_1+n_2-2}$
4. p-Value and Conclusion: Same

Two sample, Independent, Unequal variance

1. Assumptions: Same, except pop. var. is different
2. Hypothesis: $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 >< \neq \mu_2$
3. Test statistic: $T = \frac{\bar{X}-\bar{Y}}{se}$ where $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $T \sim t_{df}$ where df needs R
4. p-Value and Conclusion: Same

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.

08. Linear Regression

Simple Linear Regression

- $Y = \beta_0 + \beta_1 x + \epsilon$
- Assumptions: Random data, Relationship is linear, $\epsilon \sim N(0, \sigma^2)$ which implies $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ (Check if response var. is symmetric)
- **Ordinary Least Square Estimate** - Line with least sum of square residuals. $ss_{Res} = \sum_{i=1}^{n} e_i^2$ where $e_i = y_i - \hat{y}_i$
- **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
- If coeff. are not significant, **intercept model** $\hat{Y} = \hat{\beta_0}$

1. Assumptions: Same as building model
2. Hypothesis for T-test: $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$
3. Hypothesis for F-test: H_0 : All coeff. (except β_0) = 0 and H_1 : At least 1 coeff. is non-zero.
4. t-stat: Check from summary. Null distri: $t_{\text{n - \# of coeff}}$
5. F-stat: For simple lin. reg., $F = t^2$. Null Distri: $F_{\text{\# of coeff; n - \# of coeff}}$

Regression Diagnostics

Plots and what to look out for:

- Scatter plot $r_i vs. X_i/\hat{Y}_i$ and $X vs. \hat{Y} \to$ Linearity (Curved band), Constant variance (Funnel shape), Normality (Many points outside $(-3, 3)$)
- QQ Plot/Histogram of $r_i \to$ Normality

How to fix:

- Constant variance: Transform response variable
- Linearity: Add higher order term

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

- $|Cor(x,y)| = \sqrt{R^2} = R$
- Weakness: More variables $\to R^2 \uparrow$
- Thus, can use **Adjusted R^2**

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

09. R Code

```
matrix(c(1:6), nrow=2, ncol=3, byrow=T)
rbind(m, c(1,2,3))
data = read.csv("./crab.txt")
names(data) = c("Subject", "Gender")
data$Subject # Or attach()
data[1:8,]
data[Gender == "M" & HW == "A",]
# Replace elements based on condition
ifelse(Gender == "O", "F", "M")
# Return indices that match condition
which(flat == "3 ROOM")
for (i in 1:100) {...}
choose(6, 3)
summary(marks)
# Frequency Table
table(data)
prop.table(table(data))
# Bar Plot
barplot(table(data))
# Contingency Table
tab = table(bbd, pmh) # (r, c)
prop.table(tab) # Joint probabililty
prop.table(tab, "pmh") # Conditional
    probability on pmh groups
# Bar Plot with 2 variables
barplot(proptab)
# Boxplot
bp = boxplot(age~cancer) # quan. ~ cat.
bp$out # Values of outliers
grp = bp$group # Outliers in each group
which(grp == 1) # Outliers in group 1
bp$out[which(grp == 1)]
# Histogram
hist(flatPrice)
# Scatter Plot, Correlation
plot(size, price) # (x-axis, y-axis)
cor(size, price)
# QQ plot
qqnorm(data)
qqline(data)
# Generate vector of 10 IID samples
rbinom(n=10, size=100, prob=0.5)
rnorm(n=10, mean=100, sd=15)
rexp(n=10, rate=1/500)
# P(X >= 70)
pbinom(70, 100, 0.5, lower.tail=T)
pnorm(70, 100, 15)
# Find q0.9 where area on left = 0.9
qbinom(0.9, 100, 0.5)
qnorm(0.9, 100, 15)
qt(p=0.9, df=12)
# Generate N samples with 10 obs.
m = matrix(rnorm(10*N, 70, 10), N, 10)
sampleMeans = rowMeans(m)
# 1 Sample T-test
t.test(x=data, mu=38, alternative="less
    ", conf.level=0.95)
var.test(d1, d2) # Equal var. if >0.05
shapiro.test(d) # Test normality.
    Normal if >0.1
wilcox.test(d, 4) # If data not normal
anova(m1) # Test sig. of var. with >2
    categories. Add at end.
# 2 Sample, Independent
t.test(data1, data2, alternative="less"
    , var.equal=T, conf.level=0.95)
# 2 Sample, Dependent
t.test(d1 - d2, mu=0)
# Linear Regression
m1 = lm(price~area) # y~x
m2 = lm(price~area+type+area*type)
summary(m1)
confint(m1, level=0.95) # CI of coeff.
abline(m1) # Add fitted model to plot
# Predict
new1 = data.frame(area = c(20, 40))
predict(m1, new1)
predict(m1, new1, interval="confidence"
    , level=0.95) # with CI
rawRes = m1$res # Raw residuals
SR = rstandard(m1) # Standard residuals
which(SR > 3 | SR < -3) # Outliers
which(cooks.distance(m1)>1) #Influen.
```