ST1131 Cheatsheet

AY21/22 sem 2

github.com/Zxun2

Confounding and Lurking variables

Lurking variables influences the association between variables of primary interest.

Confounding happens when two explanatory variables are associated with the response variable and to one another. We are unable to tell which is causing the change in the response.

$\implies$ Association does not imply causation.

Probability

Additive Law of Probability

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Law of Total Probability

Suppose $B_1, B_2, \ldots, B_n$ are a partition of S. Then for any event A $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$

Bayes Theorem

Suppose $B_1, B_2, \ldots, B_n$ are a partition of S. Then for any event A,

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Epidemiological Terms

Sensitivity is the probability that the test is positive, given that the person has the disease, $Pr(pos \mid disease)$.

Specificity is the probability that the test is negative, given that the person does not have the disease, $Pr(neg| \sim disease)$

Numerical Summaries

- Mean – $\frac{1}{n}\sum_{i=1}^{n} X_i$, Variance – $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

	Discrete	Continuous
Mean	$\sum_x xp_x$	$\int xf(x)dx$
Variance	$\sum_x (x-\mu)^2 p_x$	$\int (x-\mu)^2 f(x)dx$

- Median – More representative when data is skewed
- For n **identically distributed** random variables:

$$X_1 + \cdots + X_n \sim N(\frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu, \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n})$$

- Correlation – `cor(x, y)`

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{s_x})(\frac{Y_i - \bar{Y}}{s_y})$$

QQ plots and Normality checking

Note that the sample quantiles are in the X-axis and theoretical quantiles are in Y-axis.

- Right tail above the straight line: shorter than Normal
- Right tail below the straight line: longer than Normal
- Left tail above the straight line: longer than Normal
- Left tail below the straight line: shorter than Normal

Binomial Distribution

$Bin(n, p)$ distribution is the discrete probability distribution of the number of successes in a sequence of n **independent**, each with its own **binary outcome**: success, p, or failure, 1 - p. Probability of X successes in n trials:

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

Expectation, E(X) = np Variance, Var(X) = np(1-p)

The random variable, X, takes on **n + 1** values

Can be approximated by Normal Distribution when n is large and p is not close to 0 or 1.

Normal Distribution

Normal distribution is the continuous probability distribution for a real-valued random variable. Symmetric, Bell-Shaped and Characterized by μ and σ^2.

Standardization

if $X \sim N(\mu, \sigma^2)$ and we take $a = 1/\sigma$ and $b = \mu/\sigma$, then

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

Z-scores refers to the number of standard deviations away from the mean.

Sampling Distribution

Sampling distribution is the probability distribution that specifies probabilities for the possible values that the statistic can take.

- **Central Limit Theorem** : n > 30 $\implies$ $\bar{X}$ is normal
  The approximation gets better when n increases and if X_i themselves are not too skewed
- **Data distribution** refers to the data spread in 1 sample. Gets closer to population distribution as n increases.
- The sample variance is **smaller** than the population variance.

Sample Proportion

Every X_i is $ber(1, p)$ hence, by CLT when n is large enough,

$$\hat{p} \sim N(p, \frac{p(1-p)}{n})$$

.

n should satisfy: $np(1-p) \geq 5$ for CLT to work.

Sample Mean

Case 1: Population distribution is normal

Because the X_i's are **themselves normal**, the resulting sampling distribution is also a random variable and is normally distributed.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Case 2: Population distribution is not normal

Same as case 1. Have to ensure that sample size, n, is sufficiently large $\implies$ $\bar{X}$ is still normally distributed.

Confidence Interval

A **confidence interval** contains the most believable values for a parameter.

- Formed by a method that combines the point estimate and margin of error.
- The probability that this method produces an interval that contains the parameter is called the confidence level.

$$\hat{p} \pm CI \times \sqrt{\frac{var}{n}}$$

Properties of optimal point estimates

- Should be unbiased. Has a sampling distribution that is centered at the parameter it tries to estimate.
- Smallest standard deviation compared to other point estimators.

Confidence Interval for Proportion

- Choosing the value of n given a width, D

$$n \geq (\frac{2 \times q_{1-\frac{\alpha}{2}}}{D})^2 p(1-p)$$

Confidence Interval for Mean

- Follows t-distribution with n - 1 df, `qt(quantile, df)`
- Necessary assumptions
  Normal and symmetric distribution (n < 30)
  Randomization (Not robust)
- Choosing the value of n given a width, D

$$n \geq (\frac{2t_{n-1,1-\alpha/2}(s)}{D})^2 \implies (\frac{2(q_{1-\alpha/2})(s)}{D})^2$$

Interpretation

- Long-run interpretation. 95% of such intervals will contain the parameter.
- No guarantee that the interval contain the parameter.
- All values in the interval are plausible values for population parameter.
- CI widens $\implies$ Pr(parameter lie close to estimate) $\downarrow$ (works both ways).

Notable code

- Shapiro Wilk Test for Normality – `shapiro.test(length)`

Hypothesis Testing

A **hypothesis** about a population claims that a **parameter** takes a particular numerical value.

A **Type I**, α, error occurs if we reject H0 when it is in fact true.

A **Type II**, β error occurs if we do not reject H0 when it is in fact false.

The **power** of a test is defined to be $1 - \beta$. It is the probability of correctly rejecting H0, when it is in fact false.

5 steps of Hypothesis Testing

Assumptions

- Randomization
- For categorical,
  The sample size n is sufficiently large s.t. the sampling distribution is approximately normal $\implies$ $np(1-p) \geq 5$
- For quantitative,
  Population distribution is unknown but sample distribution is slightly skewed. Ensure that $n \geq 30$.
- Variable is quantitative/categorical

State your hypotheses

- Null hypothesis, H_0
- Alternative hypothesis, H_1

Test Statistics

- For proportion. Standard error is in terms of p.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

- For mean. Standard error is in terms of s.

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}(0, 1)$$

Interpreting p-value

- Assume H_0 is true. Conditional Probability.
- A small p-value provides strong evidence against H_0

Conclusion

- Interpret the conclusion of the significance test in the context of the study. Reject H_0 $\implies$ not in the 95% CI
- Comment on the validity of H_0 when a significance level is given.

```
# for proportion, P(Z <= X)
> qnorm(X)
# for proportion, P(Z > X)
> qnorm(X, lower.tail = F)
# for mean, P(T < -X & T > X)
> t.test(data, mu = H_0,
    alternative = "two.sided",
    conf.level = 0.95)
# the interval is based on the sample "data"
# alt = c("two.sided", "less", "greater")
```

Two Sample Hypothesis Testing

Types of samples

- Independent Samples (Equal/Unequal Variance)
- Dependent Samples
  Each observation has a matched observation in the other sample.

Independent Samples

Assumptions

- Quantitative Response Variable
- Independent, randomized
- Variance is the equal/unequal
- Data is normal (crucial when n is small)

Variance Test

H_0: The two samples are from two population with the same variance – ratio is 1 `var.test(x, y)`.

Hypothesis

The null hypothesis for comparing two means has the following form:

$$H_0 : \mu_1 - \mu_2 = 0$$

The alternative hypothesis:

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Test Statistic

The **pooled estimate** of the common variance:

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

The test statistic is then:

$$T = \frac{(\bar{X}-\bar{Y})-0}{se},\ where\ se = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

For unequal variance, test statistic is:

$$T = \frac{(\bar{X}-\bar{Y})-0}{se},\ where\ se = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Interpreting p-value and Conclusion

For equal variance:
- Under H_0, T follows a T-distribution with $(n_1+n_2)-2$ degrees of freedom. `pt(t-score, df)`

For unequal variance:
- Under H_0, T follows a t-distribution with a complicated number of degrees of freedom (might not be an integer).

```
# Two sample t-test, mu1 - mu2 != 0
> t.test(count_good, count_bad,
    alternative = "two.sided",
    var.equal = T,
    conf.level = 0.95)
# Welch Two sample t-test, mu1 - mu2 != 0
> t.test(count_good, count_bad,
    alternative = "two.sided",
    var.equal = F,
    conf.level = 0.95)
```

Dependent Samples

Every observation in a sample has a matched value in other sample (Before and After).
Let $\dot{t}$ be the mean of the differences of the matched subjects in the population. Then the hypothesis is

$$H0 : \mu = 0$$

The test is then performed similar as the case of one-sample data. It is possible to use a two sample but that varies on the circumstances.

```
# Paired t-test
> t.test(Yes, No,
    alternative = "greater",
    paired = T, conf.level = 0.99)
# One sample t-test
> t.test(diff, alternative = "greater",
    conf.level = 0.99)
```

Wilcoxon Signed Rank Test

Null Hypothesis: Population median = m_0
Alternative hypothesis: Population median $\neq m_0$
- `wilcox.test(data, median)`

Linear Regression

Linear regression means that this relationship is a linear one, of the form:

$$Y = B_0 + B_1X + \epsilon$$

- ϵ is a random variable. It has a variance of σ^2 centered at zero. Variations in Y given X.

$$Y \sim N(B_0 + B_1X, \sigma^2)$$

Assumptions

- *Randomization*: Data collection
- *Relationship between X and Y is linear*: check this using a scatter plot and correlation. Add higher order terms if needed.
- Normality: Residuals, $\epsilon \sim N(0, \sigma^2)$
- *Equal Variance*: Residuals. No matter the value of X, variance is always σ^2. Transform the response: taking $\ln(Y)$, $\sqrt{Y}$ or $\frac{1}{Y}$ to be the response of model.
- The response variable should be quantitative and symmetric
  Transform the variable via log (right skewed) or exponential (left skewed).

Ordinary Least Square Estimation

Find the line that minimizes the sum of squared residuals.

$$\sum_{i=1}^{n}[e_i^2 = (y_i - \hat{y_i})^2]$$

σ is the **Residual Standard Error**.

Standardized Residuals

$$\frac{Y-\hat{Y}}{standard\ error\ of\ (Y-\hat{Y})}$$

.

Interpolation vs Extrapolation

Interpolation refers to estimating the mean response that had not been observed, but is within the range of observed values.
- A benefit of running a regression analysis.

Extrapolation refers to estimating the mean response that is outside the range of observed values.
- We do not know the form of the relationship outside our sampled values.

Confidence Band

Each sample has different variance, and different mean.
Plotting all these different samples together into a single plot creates a band in which most of these plot lines lie in.
The uncertainty around $(\bar{x}, \bar{y})$ is the smallest, but uncertainty increases as you step away from the observations.
Hence, the width of the confidence band is **not constant**.

Testing Hypothesis

There are two kinds of tests that can be conducted.
- t test: testing significance of one regressor.

$$H_0 : X_1 = 0 \quad H_1 : X_1 \neq 0$$

- F test: testing significance of whole model.

$$H_0 : \forall_i \in \{1..n\},\ X_i = 0 \quad H_1 : \exists_i \in \{1..n\},\ X_i \neq 0$$

Note that in a simple model, t-test is equal to F-test – $t = \sqrt{F}$

What plots to make?

- Plot the r_is on the y-axis against Y_i/X_i on the x-axis.
- Create a histogram of the r_i's.
- Create a QQ-plot of the r_i's.

What to look out for?

- Residual plots against Y/X should be scatter randomly about 0, within the interval (-3, 3).
- No funnel shape (Constant variance violated).
- Non-normality in QQ plot (Residuals are not normal)
- A curved band when plotting Y against X.

Standard answer: The histogram and QQ plot indicate the normality of the SR. The plot of SR versus the fitted response shows the randomness of the SR within the range of -3 to 3 with no pattern nor trend.
An **outlier** may be influential.

```
which(SR>3 |SR<(-3)) # index of outliers
C = cooks.distance(M1)
> which(C>1) # index of influential point
```

Coefficient of Determination, R^2

The proportion of total variation of the response (about the sample mean $\bar{Y}$) that is explained by the model.
In a simple model, $R^2 = cor(X,Y)^2$.
In a non-simple model, look at the adjusted R^2.

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

If there are repeated X values with different Y values, then R^2 can never be 1.

F-Test and R^2

F-test measures the significance of model while R^2 describes the variation in the response that is explained by model.

To decide between the two,
- Look at residual plots
- Compare the range of $\hat{Y}$ relative to Y

Multiple Linear Regression

Indicator Variables

An **indicator variable** takes on the value 1 if a category is observed, and 0 otherwise.

For a categorical variable with k categories, create **k - 1 indicators** each indicating 1 category. The one group for which we do not create an indicator variable will become the **reference group** for the regression.

Interaction with Categorical Variables

It is possible that one of the variable may become a linear form of another variable $\implies$ modify/drop the term

---

Useful R Code

```
# Association between cat and quant
> boxplot(quantitative~categorical)

# Linear model
> factor(categorical) # Rmb to factor
> m1 = lm(res ~ epl1 + epl2 + epl1 * epl2,
    data = dataset)

# Standardized residual
> SR = rstandard(mn)
> length(SR)
```

```
# Plot SR against Y/X
> plot(m1$fitted.values, SR,
    xlab = "Fitted values",
    main = "SR against Fv")
> abline(0, 0)

# Plot SR against categorical regressor
> plot.default(var, SR)
> abline(0, 0)

# Draw a normal curve on top of the histogram
> hist(SR, probability = TRUE, col = 2)
> x <- seq(-3, 3, length.out = 731)
```

```
> y <-dnorm(x, mean(SR), sd(SR))
> lines(x, y, col = "darkblue")

# QQ plot to test normality of residuals
> qqnorm(SR, datax = T,
    ylab = "Standardized residuals",
    xlab = "Z scores",
    main = "QQ Plot")
> qqline(SR, datax = T)

# Checking linearity, plot(x, y)
> plot(temp, cnt)
> abline(lm(cnt ~ temp,data = day))
```

```
# Generate a confint for parameters
> confint(m1, level = 0.95)

# Predict values with model
> new2 = data.frame(X1 = c(20, 30),
    X2 = c(1, 1)) # two points
> predict(m1, newdata = new2,
    interval = "confidence", level = 0.95)

# Check for insignificant regressor
> anova(m1)
```