ST1131 CheatSheet

by Zachary Chua

R Basics

Vector: set of elements of same type
`number<-c(2,4,5,6)`, c for concatenate (can be used to append)
`number<-rep(a,b)`, replicates item a, b times
`number<-seq(from=2,to=1-,by=2 (length = 5))`
Matrix: set of elements of same type in rows and columns
`v<-c(1:6)`
`m<-matrix(v, nrow=2, ncol=3, byrow=F)`, fills by column by default: 135, 246
`rbind(v1, v2)`, `cbind(m1, c1)`: to add row / col to matrix respectively
Dataframe: same as matrix but can have diff modes
Row contains diff observation, columns contain values for diff vars
Reading in csv: `data<-read.csv("file", header=TRUE)`, if not can provide vector of col names to col.names param
Access data: `data[1:4,]`, `data[Gender == "M"]`
Dataframe commands: names, attach, colMeans, which(data$col == 1) (gets index)
Vector Commands: max, min, sum, mean, range, cor(x, y), sort

Exploratory Data Analysis

Variables and Summaries
1. Quantitative: Discrete vs Continuous, check: meaningful difference
2. Categorical: Ordinal (has ordering) vs Nominal

Frequency Tables: lists all possible values, and its frequencies
- Can be expressed as a proportion or a percentage (relative frequencies)
- When summarizing, mention modal category and prop of modal category

Code: `table(data)`, `prop.table(table(data))` (for proportion)
Renaming variable: `Gender <- ifelse(Gender=="0", "Female", "Male")`
Graphical Summaries:
1. Bar Plot: display single categorical variable
   - mention groups w high or low prop, if ordinal, mention trend

Code: `barplot(table(data), ylab="", xlab="", main="", col=c(2, 5))`

2. Histograms: portray frequencies of possible outcomes of quantitative var
   - look for pattern, unimodal / bimodal, symmetric / skewed

Code: `hist(data, prob=TRUE, xlab="", ylab="", main="")`
prob=TRUE replaces frequency with density

3. Boxplot: Portrays 5 number summary of dataset, `boxplot(data)`
   - removes features like mounds / gaps
   - if dist is unimodal then gives indication of skewness
   - Report: median, outliers (how many, where), compare medians and IQR

Summary of Centre: mean and median
Median $X_{(0.5)}$: middle value, $\frac{n+1}{2}$ if odd, average of $\frac{n}{2}$ and $\frac{n}{2}+1$ if even
Mean is sensitive to extreme observations,
- for highly skewed data, median, else mean

For unimodal distributions:
1. Mean > Median: right skew
2. Mean = Median: symmetric
3. Mean < Median: Left skew

Summary of Variability
1. Range: diff betw largest and smallest (sensitive to extreme observation)
2. Variance: average of the squared deviations from the mean
$$s^2 = \frac{!}{n-1}\sum(X_i - \overline{X})^2$$
   - SD represents avg distance of observation from mean
3. IQR: distance between 75th and 25th quantiles, spread of center 50%
   - used with boxplots

Generally use var and sd with mean, IQR with median
Outlier: smaller than $Q_1 - 1.5 \times IQR$ or larger than $Q_3 + 1.5 \times IQR$
Association between 2 Vars: response and explanatory variable
Response: variable on which comparisons are made
Explanatory: variable you believe response depends on
1. **2 Categorical Variables**
1.1 Contingency Table: row — explanatory, col — response
   - Each entry is the number of observations (or relative proportions)
   - Proportions can be row-wise, column-wise or joint
   - Relative Risk: ratio of percentages, resp and exp / resp and not exp
   - If v diff from 1 can show association

Code: `table(row, col)`, `prop.table(table, "name")`
   - if name is row, then row-wise proportions

1.2 Barplots: clustered / stacked
Code (clustered): set `beside=TRUE`, default is stacked
2. **1 Categorical, 1 Quantitative**: side by side boxplot
Code: `boxplot(var1 var2)`, `plot$out` for outliers
3. **2 Quantitative**
3.1 Scatterplot
   - Relationship? +ve / -ve / no assoc
   - can be approx by straight line?, how do points vary about line? outliers?

Code: `plot(var1, var2)`
3.2 Correlation: r, always between -1 and 1, measures linear association
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right)$$
Code: `cor(v1, v2)`

Data Collection

Lurking Var: unobserved, influences assoc. between variables of interest
Confounding: 2 explanatory variables are assoc. with response var, but also with each other, hence cannot tell which is causing the change in response
- Lurking variable has potential for confounding

Sampling Survey for Observational Studies
1. Identify Population
2. Create sampling frame
   2.1 Sampling frame: list of subjects in popn from which sample is taken
3. Specify method for choosing subjects from 2, aka sampling method
4. Collect data from sample

Simple Random Sample: each possible sample of that size has same chance of being selected
How: Subjects numbered, generate n random numbers, subjects with those numbers are chosen
Data Collection
1. F2F interview: easier to get people to respond, but costly
2. Telephone: cheaper, but easier for people to refuse
3. Self-administered questions: cheaper, less labour, but lower response rate

Bias in Sample Surveys
1. Sampling bias: not random, or sampling frame not representative
2. Nonresponse Bias: sampled subjects unreachable / refuse to participate
3. Response Bias: not honest / answer wrongly, may be due to how question asked / phrased

Bad Sampling: Convenience / Volunteer samples
Experimental Studies: Controlled, random, blind
- Randomisation: eliminate bias, balance out lurking variables,

Random Variables and Probability

Probability:
1. Sensitivity: test positive, given that person has disease
2. Specificity: test is negative, given that person does not have the disease

Properties of Mean (discrete)
1. $E(\overline{X}) = \frac{1}{n}\sum X_i = \mu$, expectation of sample mean is mean

Variance (discrete): $\sigma^2 = \sum_x (x-\mu)^2 p_x$
Properties:
1. $Var(\overline{X}) = \frac{1}{n^2}\sum \sigma^2 = \frac{\sigma^2}{n}$, variance of sample mean

Quantiles: 100p-th quantile, q_p, such that $P(X \leq q_p) = p$
Poisson Approximation for Binomial
- Large n, small p can be approx. by Poisson with param np

Properties of Normal Distribution
1. Adding constant to normal is still normal
2. Sum of normals is normal
   2.1 $X + a \sim N(a + \mu_X, \sigma_X^2)$
   2.2 $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
   2.3 $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$
3. Product of normal with constant is normal
   3.1 $aX \sim N(a\mu_X, a^2\sigma_X^2)$

Standardisation: $\frac{X-\mu}{\sigma}$, Z-score of X
Normal Approx for Binomial
- n large, p not too extreme $np(1-p) \geq 5$, can be approx by $N(np, np(1-p))$

R: for normal distribution
- Generate vector of 6: `rnorm(6, mean=100, sd=15)`
- $P(X \leq 115)$: `pnorm(115, 100, 15, lower.tail=TRUE)`
- $q_0.9$: `qnorm(0.9, 170, 10)`

Sampling Distribution

Definition: Prob dist of a statistic (prob for value of a statistic)
Central Limit Theorem:
Suppose iid $X_1, \ldots, X_n$, $n \geq 30$ for quant, $np(1-p) \geq 5$ for cat, then sample mean/proportion approx $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
Sample Distribution of sample proportion, $\hat{p}$
Sample proportion, $\hat{p} = \frac{1}{n}\sum_{i=0}^{n} X_i$
Sampling Distribution of $\hat{p} \sim N(p, \frac{p(1-p)}{n})$ approximately if $np(1-p) \geq 5$
Since we do not know p, estimate with $\hat{p}$, sd is estimated to be
$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ standard error of } \hat{p}$$
Sampling Distribution of sample mean (normal)
If indiv X_is are normal, then $\overline{X} \sim N(\mu, \sigma^2/n)$ (exactly, not approx)
Sampling Distribution of sample mean (not normal)
$\overline{X} \sim N(\mu, \sigma^2/n)$, approximately if $n/geq30$
Observations:
1. Variability decreases as n increases
2. bell shapes are all centered at pop mean
3. sampling distn of $\overline{X}$ depends on μ, σ^2, n (not N)

Data Distn: histogram from one sample, if n large, resemble popn dist
Note: s is sd of sample, $s/\sqrt{n}$ is estimated sd of sample mean, aka standard error

Confidence Intervals

A type of statistical inference

Point Estimate: single number that is best guess of pop param
1. for mean μ, use sample mean $\overline{X}$
2. for proportion p, use sample proportion $\hat{p}$

Note: change for each sample, no idea how close to actual param
Properties of Optimal Point Estimate:
1. Unbiased, sampling dist should be centered around μ
2. Small variances

Interval Estimate: interval within which the param is believed to fall
point estimate $\pm$ margin of error (multiple of sd of sampling dist)
Confidence Intercal for Population Proportion
Formula: $\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$, 1.96 for 95% CI
Estimate p with $\hat{p}$, becomes standard error (used when $np(1-p) \geq 5$)
Procedure with 100x CI:
1. Get $\hat{p}$, ensure $n\hat{p}(1-\hat{p}) \geq 5$, or increase n
2. Find $\alpha = 1 - x$
3. Find value of $q_{1-\alpha/2}$ from $N(0,1)$
4. CI $= \hat{p} \pm q_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Determining Sample Size: $n \geq (\frac{2 \times q_{1-\alpha/2}}{D})^2 p(1-p)$, using $p = 0.5$
- for a certain CI with width D or less

Confidence Interval for Population Mean

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}, \text{ not normal}$$

Get Quantile values from R: `qt(quantile, df)`
Formula: $\overline{X} \pm t_{n-1,0.975} \times \frac{s}{\sqrt{n}}$
Procedure for 100x CI:
1. find $\overline{X}$ from sample, find $\alpha = 1 - x$
2. Derive $t_{n-1,1-\alpha/2}$
3. CI $= \overline{X} \pm t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$

Determine sample size: $n \geq (\frac{2t_{n-1,1-\alpha/2}s}{D})^2$
- approx t dist with N(0,1)
- approx s by looking for similar study or pilot study

Interpreting CI: Confidence refers to long-run interpretation
Describes how well method performs over many different random samples (95% of them will contain pop param)
Note: Given a particular 95% interval, cannot tell whether it contains the true pop param
CI and sample size: Width increases when reduce sample size

Hypothesis Testing

5 steps of Hypothesis Testing
1. Assumptions: data from randomisation, sample size, pop dist (normal?)
2. State Hypothesis: H_0 (no effect) and H_1 (some effect)
   2.1 Test side: $\neq$: two-sided, $>$: right-sided, $<$: left sided
3. Test Statistic: Distance in number of se from point estimate to H_0
   3.1 need point estimate, sampling dist (null dist), value under H_0
4. p-value: Prob of value that or more extreme, assuming H_0 is true
   - small p-value, strong evidence against H_0
5. Conclusion, reject or retain H_0, comparing to sig level

Hypothesis Testing for Proportions
3. Test Statistic: $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, $Z \sim N(0,1)$
4. p-value, eg $Z = 3.866$
   4.1 two-sided: pvalue is two areas (left and right tail)
   Code: `2*pnorm(3.866, lower.tail=FALSE)`
   4.2 right-sided: pvalue is right area of test statistic
   Code: `pnorm(3.866, lower.tail=FALSE)`
   4.3 left-sided: pvalue is left area of test statistic
   Code: `pnorm(-3.866)`

Hypothesis Testing for Means
Same except that Test Statistic T follows t dist with $n-1$ df
T-test in r:
`t.test(data, mu=mean, alternative="two-sided", conf.level=0.95)`
alternative can be "less" or "greater"
Errors
1. Type 1: reject when it is true, probability is α
2. Type 2: do not reject when it is false, probability is β
   2.1 power of a test is $1-\beta$, prob of correctly rejecting

Cannot reduce both simultaneously
Shapiro test: tests normality (large p $\rightarrow$ normal)
Wilcoxon test: for when not normal, tests median
2 sample Hypothesis Testing
Independent Sample, Equal Variance
1. Assumptions: quantitative, 2 indep samples, variance is same, normal dist
   1.1 Test for equal variance using `var.test(x, y)`
2. Hypothesis: $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2, \mu_1 < \mu_2, \mu_1 > \mu_2$
3. Test Statistic: Used pooled estimate of variance
   3.1 $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
   3.2 $se = s_p\sqrt{1/n_1 + 1/n_2}$
   3.3 $T = \frac{(\overline{X}-\overline{Y})-0}{se}$
4. T follows t dist with $(n_1+n_2)-2$ df

Code: `t.test(mu1, mu2, alternative="two.sided", var.equal=TRUE, conf.level=0.95)`
Independent Samples, Unequal Variance
3. Test Statistic: $T = \frac{(\overline{X}-\overline{Y})-0}{se}, se = \sqrt{s_1^2/n_1 + s_2^2/n_2}$
   3.1 T follows t dist with complicated df (let R calc)

Code: `t.test(mu1, mu2, alternative="greater", var.equal=FALSE, conf.level=0.99)`
Dependent Samples: each observation has matched observation in other sample
Treat as set of differences, let μ be mean of differences, $H_0 : \mu = 0$
Code: `t.test(diff, mu=0, alternative="Greater", conf.level=0.99)`
Code: `t.test(mu1, mu2, alternative="greater", paired=TRUE, conf.level=0.99)`
These two lines are equivalent

Linear Regression

Regression: mathematical relationship between mean of response Y and diff values of X
Linear Regression: $Y = \beta_0 + \beta_1 X + \epsilon$
1. $\epsilon \sim N(0,1)$, represents error
2. β_0 is y-intercept, β_1 is slope of line, are pop params
3. Linear refers to linearity in parameters, simple means 1 explanatory

Assumptions
1. Data obtained by randomization
2. Relationship betw X and Y linear
3. Assume $\epsilon \sim N(0,1)$
4. Equal Variance

Check 2-4 after model fitted
Specifications
1. For any X, resp var observed has normal dist $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$
2. For any X, mean of resp var is $\beta_0 + \beta_1 X$
3. Whatever X, variance of resp var is always same σ^2

Code: `M1 = lm(resp exp, data=dataset)`, then `summary(M1)`
Fitted Model: model without ϵ with values subbed in for β_0 and β_1
eg. $\hat{Y} = 0.5 + 12.67X$, then $\hat{Y}$ is point est of mean of response at X value
Fitted values: `new1=data.frame(exp=c(20,30)); predict(M1, newdata=new1)`
Note: Can only interpolate, cannot extrapolate
Estimating σ^2: point estimate from residuals. $e_i = Y_i - \hat{Y}_i$
- $\hat{\sigma}$ = Residual Standard error in R output

Interval est for β_0 and β_1: `confint(M1, level=0.95)`, var name optional
Interval est for mean response:
`predict(M1, newdata=new1, interval="confidence", level=0.95)`
Significance tests
1. t test for one regressor:
   1.1 Assumptions same as model
   1.2 $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$, one sided also can
   1.3 Test statistic is t-statistic from R output, p-value also from R output
   1.4 Null dist: t dist, df $= n-2$, n is number of coeff and intercept
2. F test for whole model
   2.1 H_0 : all coeff except intercept, are 0, H_1: at least one coeff is non-zero
   2.2 Test statistic and p-value found in R output.
   2.3 F-test not significant, $\hat{Y} = \hat{\beta}_0$, Y doesnt depend on any regressor
      2.3.1 Then put 1 for explanatory variable in `lm`

Regresssion Diagnostics: Checks if adequacy of model
1. Linearity: check using scatter plot betw X and Y and residual plot
   1.1 Fix: use higher order terms in X
2. Normality: checked using residuals
3. Equal Variance: residuals
   3.1 Fix: transform response (ln, sqrt, reciprocal)

Getting Residuals: `M1$res` (raw res), `rstandard(M1)` (Standardised res)
Plots with residuals:
1. Plot r_is on y-axis against $\hat{Y}_i/\hat{X}_i$ on x-axis
   1.1 Expect scatter randomly about 0 (not funnel shaped), within (-3, 3)
2. Histogram, expect normal
3. QQ-plot, expect normal

Outlier: SRs > 3 or < -3
Code (index): `which(SR > 3 | SR < (-3))`
Influential Point: affects param value alot, Cook's distance (> 1)
Code: (C = cooks.distance(M1)) and `which(C > 1)`
Coefficient of Determination, R^2: Checks goodness of fit
Measures how much of the variation of the resp can be explained by model
In simple model, correlation coeff $= \sqrt{R^2}$, if $\hat{\beta}_1 > 0$ and vice versa
Adding more variables increases R^2, but increases complexity of model
- use adjusted $R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$, where k is no. of variables in model

Categorical Variables, and Indicator Variables
Add categorical variables to model using indicator variables
Interaction terms: if there is interaction betw 2 variables, add product of them to model
Note: if variable is not significant, but interaction term containing var is, need to retain the variable

QQ Plots

For sample quantiles on X-axis, and theoretical quantiles on Y-axis,
1. R tail below / above line: longer / shorter than Normal
2. L tail below / above line: shorter / longer than Normal

Opposite if X and Y axis swapped
Code
```
qqnorm(SR, datax=TRUE, ylab="SR", xlab="Z scores", main="")
qqline(SR, datax=TRUE, col="red")
```