ST1232 cheat sheet

Statistics for Life Sciences (National University of Singapore)

Hypothesis Testing

1. Assumptions:
   - Data comes from randomization
   - Shape of population distribution is normal
   - Sample size large enough
2. Hypothesis:
   - H_0: null hypothesis
   - H_1: parameter falls in alternative range of values
3. Test statistic:
   - How far point estimate falls from the parameter value given in null hypothesis
   - Measured by number of standard errors between point estimate and parameter value in H_0
4. Interpreting test statistic:
   - Presume H_0 is true
   - But if sample test statistic falls out in the tail of sampling distribution, value is unlikely
   - P-value is the probability of "what we saw or something more extreme" given H_0 is true

Significance level:

- a number such that we reject H_0 if p-value is less than or equal to that number
- if we reject H_0, the results are statistically different

Hypothesis Testing for proportions

1. Assumptions:
   - Variable collected is categorical
   - Data comes from randomization
   - Sample size n is sufficiently large that the sampling distribution of sample proportion $\hat{p}$ is approximately normal if $np_0(1-p_0) \geq 5$
2. Hypothesis
   - H_0: $p = p_0$
   - H_1: $p \neq p_0$
3. Test statistic:
   - $\hat{p}$ N(p_0 , $\frac{p0(1-p_0)}{n}$) when H_0 is true
   - $$Z=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$, Z $(0,1)$

Hypothesis testing for mean (One sample t-test)

1. Assumptions:
   - Variable collected is quantitative
   - Data comes from randomization
   - Population distribution is approximately normal
   - n is small → one sided t-test
2. Hypothesis:
   - H_0: $\mu=\mu_0$
3. Test statistic:
   - Distance between sample mean $\acute{X}$ and H_0 value of population mean
   - Measured in terms of standard errors
   - $$T=\frac{\acute{X}-\mu_0}{s/\sqrt{n}}$$ with n-1 degrees of freedom
4. Conclusion:
   - If H_0 is true
   - Mean is not significantly different from μ_0

Normality assumption

1. Histogram with a Normal pdf overlaid
   - Symmetric, unimodal
2. Q-Q plot
   - Standardized sample quantiles against theoretical quantiles of N(0,1) distribution
   - If fall on a straight line, the data is normal
   - Right tail below straight line: fatter
   - Left tail below straight line: thinner

Errors in conclusions

1. Type 1 error: occurs if we reject H_0 when in fact it is true
   - when H_0 is assumed to be true
2. Type 2 error: occurs if we do not reject H_0 when it is false
   - When H_1 is assumed to be true
3. ↑sample size → type 2↓
4. ↓significance level, type 1↓ but type 2↑

2 independent sample t-test, Equal variance

1. Assumptions:
   - Independent samples with randomized experiment
   - Sample standard deviations not twice…
   - Population distribution is approximately normal and n is small
2. Hypothesis:
   - H_0: $\mu_1=\mu_2$
3. Test statistic:
   - Point estimate of difference between the population means $\acute{X}-\acute{Y}$
   - Estimate of common variance: s_p^2 = $$\frac{(n_1-1)s_1^2+(n_2-}{n_1+n_2-2}$$
   - $$T=\frac{(\acute{X}-\acute{Y})-0}{se}$$ with n_1+n_2-2 degrees of freedom
   - se= $s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

2 independent sample t-test, Unequal variance

1. Assumptions:
   - Quantitative variable for 2 groups, independent, random
   - Sample standard deviation twice…
   - Approximately normal and n is small
2. Hypothesis:
   - H_0: $\mu_1=\mu_2$
3. Test statistic:
   - Point estimate of difference between the population means $\acute{X}-\acute{Y}$
   - $$T=\frac{(\acute{X}-\acute{Y})-0}{se}$$ with complicated number of degrees of freedom (df)
   - se= $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
4. Conclusion:
   - If do not reject H_0, no statistically difference between means

Dependent sample test

-Testing if $\mu_1=\mu_2$

1.Construct new variable

$-D_i=X_i-Y_i$

2.Hypothesis:

$-H_0$: $\mu_D=0$

3.Test statistic:

- $$T=\frac{\acute{X}-\mu_0}{s/\sqrt{n}}$$

with n-1 degrees of freedom

4.Pairing up reduced standard error in $\acute{D}$

ANOVA: Compare more than 2 Group means

1. Assumptions:
   - Population distribution for the response for k groups are normal
   - Largest group sd less twice of smallest sd
   - Data randomized
   - N observations in total: n observations in k group, N=nk
2. Hypothesis:
   - H_0: $$\mu_1=\mu_2=\mu_k$$
   - H_1: At least 2 means are not equal
3. Test statistic:
   - Involves within grp &

Multiple comparison test

-To control Experiment error rate **a**: the probability of making at least 1 Type 1 error when all m hypothesis are true

1. Bonferroni correction:

- Not doing ANOVA
- perform each test at a/m significance level
- for each CI at (1-a/m)100% level
- p-value ¿ $\frac{a}{m}$

2.Tukey:

- replacing use of t-distribution for each pairwise distribution with a multiplier from another

Wilcoxon signed rank For matched pairs

1. Assumptions:
   - 2 groups are independent
   - Data random
   - Difference of observations can be ranked
   - Population distribution of difference is symmetric
2. Hypothesis:
   - H_0: $q_{0.5}=0$
3. Test statistic:
   - Sum of the ranks of the difference that were

2-sample Wilcoxon rank sum (Mann-Whitney U)

1. Assumptions:
   - Independent observations $X_1, X_2…X_{n1}$ and $Y_1,Y_2…Y_{n2}$ from 2 groups
   - Random data

Hypothesis:

- H_0: 2 samples are from same distribution

2. Test statistic:
   - Minimum of R_x and R_y, R_x …e sum of …anks of…

Kruskal-Wallis

-If more than 2 groups of independent samples

-Not normal and unequal variance

1. Assume:

-Independent observations from k groups where k ¿ 2

-Random data

2.Hypothesis:

-H_0: The distributions of all groups …the same

Simple linear Regression

-F-test equals t-test

1. Assumptions:
   - Random data
   - R/S between X and Y is linear
   - $\varepsilon\ N(0,\sigma^2)$, normally distributed subpopulation for each explanatory variable and same variance
   - $Y\ N(\beta_0+\beta_1X$
2. t-test Hypothesis:

Multiple linear Regression

-more than one explanatory variable

-t-test for individual coefficient(quantitative)

-F-test for overall significance (categorical)

-Use adjusted R^2 to compare models

Indicator variables

$$\hat{Y}=\beta_0+\beta_1X_1+\beta_2$$

when reference=female (0) &male (1)

1. Hypothesis:
   - H_0: all coefficients=0
   - H_1: at least 1 not zero

Point estimates

- $\hat{\beta}_0$ (constant) and $\hat{\beta}_1$ are point estimates
- $\hat{Y}$ is a point estimate of the mean of response at a particular X value

-when X increases by 1 unit, mean response $\hat{Y}$ increases by $\hat{\beta}_1$ unit

Compare within 2 groups

-H_0: $\mu_i = \mu_j$ if F-test is significant

- $$T = \frac{\acute{Y}_i - \acute{Y}_j}{se}$$

where se= $s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ with N-k degrees of freedom

between grp variability

- F= $\frac{Btwn\ grp\ vari}{within\ grp\ vari}$ with numerator k-1 and denominator N-k degrees of freedom

4. p-value: always the areas to the right of it; larger F→ against H_0

distribution (a studentized range distribution)

- input: no. of k grps; output: all significant pairwise comparisons

3. Dunnett:

-k groups with one control, conduct m=k-1 comparisons

-same as Tukey but will provide shorter CI

positive (W+)

- If W+ ≈ W-, do not reject H_0

For one sample

1. Hypothesis:
   - H_0: $q_{0.5} = m_0$
2. Test statistic:
   -sum of ranks $¿ m_0$ known as (W+)

sample while R_Y is the sum of ranks of Y-sample

- H_0 is correct if $R_X \approx R_Y$

3.Test stats:

-Uses the variability of sample-mean ranks

-Under H_0, follows a X^2_{k-1} distribution (chi-square)

- H_0: $\beta_1 = 0$ and t-distribution follows n-2 degrees of freedom

3. F-test Hypothesis:
   - H_0: $\beta_1 = 0$ and F distribution with 1 and n-2 degrees of freedom

2. **Adjusted** R^2 =0.068 means only 6.8% of variation in the marks change is explained by regression model

With interaction term

$$\hat{Y} = \beta_1 + \beta_1 X_1 + \beta_2$$

Estimating σ^2

-how variable the response measurements are around fitted line

- $\widehat{\sigma}^2$ is computed using residuals

$$e_i = Y_i - \hat{Y}$$

-Variance is minimized at $X_0 = \acute{X}$, smaller the better

Scatterplot

1. Linearity violated:
   - Add higher order terms in X eg. X^2
2. Variance not constant:
   - Transform the response by taking ln(Y), $\sqrt{Y}$ or (1/Y)
   - 1 unit increase in X, mean response $\hat{Y}$ increase by $e^{\widehat{\beta}_1}$ times

Residuals plot

- Detect non-normality
- Check for non-constant variance and the need to transform Y
- Check for the need to add higher order terms in X

Explanatory data analysis

1.Frequency table:

-For categorical or quantitative variable

-Modal category: highest frequency

Plots to make

1. Plot r_i's against $\hat{Y}_i$ on
2. Plot r_i's against X_i
3. Create histogram of r_i
4. Create QQ plot using r_i

- Transform Y when there is funnel shape
- Add higher order X term when there is a curve band
- Non-normal when points are outside 3 and -3

*Note that r_i are not independent

Cook's distance plot

- To check for influential point

R^2: The proportion of total variation of the response that is explained by the fitted regression.

4.Boxplot

-Minimum, lower quartile, median, upper quartile and maximum

-For comparing between quantitative vs categorical variable

-Skewness of distributions (mention if obvious)

Chi-square test for 2 × 2 with continuity correction

1. Assumptions:
   - 2 categorical variables
   - all expected cell count >5
   - Data random
2. Hypothesis:
   - H_0: 2 Variables are independent
   - H_1: 2 Variables are dependent
3. Test statistic:
   - $$\chi^2 = \sum \frac{(|observed\ cou}{ex}$$
4. P-value:
   - Right tail probability of χ^2 distribution with 1 degree of freedom

Pearson chi-square

- r rows and c columns that define 2 categorical random variables
- use $\chi^2_{(r-1)(c-1)}$ distribution

1.Mean

Fisher's exact

1. Assume:
   - ≥ 20% expected cell count <5
   - 2 binary categorical variables
   - Data random
2. Hypothesis:
   - H_0: 2 variables are independent
3. Test stats:
   - The first cell count

Linear by linear

-There is at least 1 ordinal variable

6.pth-quantile

Strength of association

-How different is P(disease|exposure) is from P(disease|no exposure)

- $\widehat{p}_1 = \frac{a}{a+b}$ is point estimate of P(disease|exposure)
- $\widehat{p}_2 = \frac{c}{c+d}$ is a point estimate of P(disease|no exposure)

Prospective study:

-randomly assign the exposure variable to subjects or record their exposure variable status

Retrospective study:

-case control study

-cannot obtain $\widehat{p}_1$ and $\widehat{p}_2$ from 2 × 2 table

Probability

Difference of proportion:

- $\widehat{p}_1 - \widehat{p}_2$

Relative risk:

- $\frac{\widehat{p}_1}{\widehat{p}_2}$

Odds of success:

-odds= $\frac{p}{p-1}$ where p is the probability of success and p-1 is failure

-odds ratio= $\frac{a/b}{c/d} = \frac{ad}{bc}$

-The odds of having disease in expose is n times the odds of having disease in unexposed

Computing CI for OR

-estimated se for ln(OR) is se(ln(OR))= $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

-Formula for a (1- α)100% CI:

$$e^{\ln\left(\frac{ad}{bc}\right) \pm q_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+}}$$

Random variable:

Logistic regression

-Response variable Y is a binary random variable (categorical)

1.Assumptions:

-Data is randomized

-Observations are independent

-Denote p_i=P(Y_i=1) (success probability) given a value of X_i

2.Odds of success:

- $$\frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{p_i}{1 - p_i}$$

- ln ($\frac{p_i}{1-p_i}$)= $\beta_0 + \beta_1 X_i$

- ln ($\frac{p_i}{1-p_i}$)= $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + .$

3.Distributional assumption:

- $Y_i \sim Bin(1, p_i)$
- $$p_i = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

Maximum likelihood estimation:

-Test fitness of model using R^2

Binomial Distribution

Model interpretation

-For every increase in X_i, the ln odds increase by β_1 times

-For every increase in X_i, the odds of success will increase by e^{β_1} times

Wald Hypothesis Test

-Test if an individual term is significantly different from zero

1. Hypothesis:

-H_0: $\beta_1 = 0$

2. Test statistic:

- $$\frac{1}{\widehat{\beta}_{¿}} \quad s.e(¿¿)^2 \text{ with } \frac{\widehat{\beta}_1}{¿ ¿}$$

sampling distribution χ^2_1 distribution

Confidence Interval of Wald Test

- $$\widehat{\beta}_1 \pm q_{1-\frac{\alpha}{2}} \times s.e(\widehat{\beta}_1)$$

-odds ratio corresponding to β is ($e^L, e^U ¿$

-95% CI for odds ratio corresponding to 1 unit increase in X_i is ($e^L, e^U ¿$

Confidence interval

-Look at proportion of modal category, compare it via the difference

2.Bar Plot:

-To display single categorical variable

-Mention group of categories with high/low proportions

-Mention any apparent trend in proportions if there is ordering

3.Histogram

-use bars to portray the frequency or relative frequency of the possible outcomes for quantitative variable

-Overall pattern: Gap, outlier?

-Skewed or symmetric?

-Unimodal? Bimodal?

-Spread of data

-Median and IQR

-No. of outliers

-Outlier: 1.5 $\times IQR$ below and above upper/lower quartile

5.Contingency table

-To display 2 categorical variables

6.Scatterplot

-For association between quantitative vs quantitative variable

-correlation:

$$\frac{Y_i-\acute{Y}}{s_y}$$

$$\left(\frac{X_i-\acute{X}}{s_x}\right)(¿)$$

$$r=\frac{1}{n-1}\sum_{i=1}^{n} ¿$$

-r is between 1 and -1

-Sum of observations divided by number of observations

- $\acute{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$

2.Median

-Middle value of the observations when observations ordered from smallest to largest

- $\left(\frac{n+1}{2}\right)$ th largest if odd, average of $\left(\frac{n}{2}\right)$ th and $\left(\frac{n}{2}+1\right)$ th if even → $X_{0.5}$

*If date is highly skewed, report median and if symmetric, report mean

3.Range

-Difference between largest and smallest observations

-Measure of spread but sensitive to extreme observations

4.Variance

-Average of the squared deviation from the mean

- $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\acute{X})^2$

5.Standard deviation

-Represents average distance from mean

-68% observations fall between 1 sd of mean, between $\acute{X}-s$ and $\acute{X}+s$

-95% fall between 2 sd of mean

-All fall between 3 sd of mean

-value that p percent of the observations fall below or at that value ($\widehat{q_{0.5}}=X_{0.5}¿$

7.IQR

-Distance between upper and lower quartile

$$(\widehat{q_{0.75}}-\widehat{q_{0.25}})$$

-50% of sample points fall within IQR

-how spread the 'middle' data is

Lurking variable

-Variable that is usually unobserved that influence the association between variables of primary interest

Confounding

-2 explanatory variables are associated with a response variable but also associated with each other

Poor sources of bias

-Sampling bias

-Non-response bias

-Response bias

1.Mutually exclusive (disjoint)

-A and B has no common outcome

$$A\cap B=\varnothing$$

-If P(A) ≥ 0 , P(S)=1,

$$P(A\cup B)=P(A)$$

2.Not mutually exclusive

- $P(A\cup B)=P(A)$

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

3.Independent event

-2 events do not influence one another

- $P(A\cap B)=P(A)$

- $P(A|B)=P(A)$

4.Conditional probability

-Probability of A given B

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

5.Bayes Theorem

6.Sensitivity

-Probability that the test is positive given that the person has disease P(A|B)

7.Specificity

-Probability that the test is negative given no disease $P(A^c|B^c)$

8.Prevalance

-Having Disease P(B)

-Numerical measurement of outcome of experiment via random sampling

$$p_z=\frac{e^{-20}20^z}{z!}$$

for barplot

1.Mean:

-sum of probabilities multiplied by possibilities

$$\mu=\sum_x x\,p_x$$

2.Expected value of X, E(X): The mean of the probability distribution of a random variable X

$$E(\acute{X})=\frac{1}{n}\sum_{i=1}^{n}X_i=$$

*mean of variables $\acute{X}$ =mean of each random variable

3.Standard deviation

-measures variability of a random variable from the mean

$$\sigma^2=\sum_x (X-\mu)^2 p$$

4.Variance

-large standard deviation has larger variability

$$Var(\acute{X})=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2$$

Probability density function pdf

-curve that determines probabilities of intervals

Mean of continuous variable X

$$\mu=\int xf(x)dx$$

Variance of continuous variable

$$\sigma^2=\int (x-\mu)^2 f(x$$

Pth quantile of cont variable

$$P(x\le q_p)=p$$

Combination:

$$C_k^n=\frac{n!}{k!(n-k)!}=$$

-n trials have 2 possible outcomes, independent

-same probability of success

$$X\ \ Bin(n,p)$$

$$P(X=x)=C_x^n p^x(1-$$

$$1-p$$

$$E(X)=np, Var(X)=$$

)

Normal distribution

-Symmetric, bell-shaped

$$X\ \ N(\mu,\sigma^2)$$

$$\acute{X}=\frac{X_1+..X_n}{n}\ \ N(\mu$$

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

Probability distribution

P(X=1)=p, P(X=0)=1-p

Sampling distribution:

-Probability distribution that specifies the probabilities for possible values that statistic can take

Sampling proportion $\hat{p}$ for random sample size n:

Mean=p, sd= $\sqrt{\frac{p(1-p)}{n}}$,

$p,\left(\sqrt{\frac{p(1-p)}{n}}\right)^2$ if

$\hat{p}\ \ N¿$

n is sufficiently large that $n\hat{p}(1-\hat{p})\ge5$, the sampling distribution is approximately normal

CLT: $\acute{X}\ \ N\left(\mu,\frac{\sigma^2}{n}\right)$

-independent observations

-n $\geq$ 30

-X is approximately normal

*a long run interpretation

Margin of error:

-measures how accurate the point estimate is likely to be estimating a parameter→sd

Standard error:

-an estimated deviation of a sampling distribution

Population proportion: p

Sample proportion: $\hat{p}$

→for cat

When $n\hat{p}(1-\hat{p})\ge5$,

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

99%=2.58,80%=1.28,95%=1.96

Width: higher the CI, smaller α and wider the width

$$2\times q_{1-\frac{\alpha}{2}}\times\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n=\left(\frac{2\times q_{1-\frac{\alpha}{2}}}{D}\right)p(1-$$

*use p=0.5 (highest variability to obtain highest n)

t-distribution:

-probability on the degree of freedom(df)

-thicker tail, more variability than N(0,1)

-larger the df, closer to N(0,1)

$$\acute{X}\pm t_{n-1,1-\frac{\alpha}{2}}\times\frac{s}{\sqrt{n}}$$

$$2\times t_{n-1,1-\frac{\alpha}{2}}\times\frac{s}{\sqrt{n}}\le D$$

$$n\ge\left(\frac{t_{n-1,1-\frac{\alpha}{2}}\times s}{D}\right)^2$$