Advanced Statistics

Lecture 1 and 2

* Independent, Identically Distributed (IID)
* Bernoulli trials
* Permutations : $\frac{n!}{(n-r)!}$ where n is total things, r is what we need to choose from n, no repetetion, order doesn't matter.
* Combinations : $\frac{n!}{r!(n-r)!}$
* Geometric Distribution : $\mu = \frac{1}{p}$, $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{q}{p^2}}$
  $\therefore q = 1-p$
* Geometric probability : $P(\text{success on } n^{th} \text{ trial}) = (1-p)^{n-1} p = q^{n-1} p$
* Binomial probability : $P(\text{K successes in } n \text{ trials}) = {}^{n}C_K \, p^K q^{n-K}$

where 'K' is number of successes in 'n' trials.

* Binomial Distribution : $\mu = np$, $\sigma = \sqrt{npq}$

Lecture 3

* Poisson probability : $P(\text{observe K rare events}) = \frac{e^{-\mu} \mu^K}{K!}$

where $K = 0, 1, 2, \cdots$, $\mu = \lambda t$

* Poisson Distribution : $\mu = \lambda t$ where 'λ' is average number of outcomes & 't' is time or space. ** $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
* Negative Binomial : $P(K^{th} \text{ success on } n^{th} \text{ trial}) = {}^{n-1}C_{K-1} \, p^K q^{n-K}$
* Normal Distribution : amount of data that falls onto different regions of a normal dist. :

$\mu \pm \sigma \rightarrow 68\%$ data, $\mu \pm 2\sigma \rightarrow 95\%$ data

$\mu \pm 3\sigma \rightarrow 99.7\%$ data

Lecture 4

* Point Estimates and Sampling Variability (concepts)
* Margin of error (concept)
* Central limit theorem (CLT) : $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$

Lecture 5

* CLT conditions : ① Independence ② Sample size

  $SE = \sqrt{\frac{pq}{n}}$ only when np & nq are atleast 10. (when p is known)
* Confidence interval : (95%) : point est. $\pm$ 1.96 $\times$ SE

  (in general) : point estimate $\pm$ z $\times$ SE
* Width of an interval (if too wide, it is not very informative)
* margin of error : z $\times$ SE = ME
* z-values for certain confidence levels

z-score	p-value	confidence level
$\pm$ 1.65	< 0.10	90%
$\pm$ 1.96	< 0.05	95%
$\pm$ 2.58	< 0.01	99%
$\pm$ 2.33	< 0.01	98%

* Interpretation of confidence intervals
* Difference b/w SD (Standard deviation) and SE (Standard error)

Lecture 6

* Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

where $P(A \cap B) = \frac{n(A \cap B)}{n(B)}$ for general $P(B)$

or

$P(A \cap B) = P(A)P(B)$ for independent events

$P(A \cap B) = 0$ for mutually exclusive events

* Law of total probability: $P(E) = P(E|F)P(F) + P(E|F^c)P(F^c)$
* Bayes Theorem: $P(F|E) = \frac{P(E|F)P(F)}{P(E)}$
* Hypothesis Testing: using confidence intervals
* Decision errors (confusion matrix)

		Prediction	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type error 1
	H_A true	Type error 2	✓

* Type 1 error is rejecting H_0 when it was actually true.
* Type 2 error is accepting H_0 when it was actually false.
* Significance level

Lecture 7 and 8

* Hypothesis testing: (Steps)

① Set hypothesis: • H_0: μ = null value
• H_A: $\mu <$ or $>$ or $\neq$ nullvalue

② Calculate point estimate: $\hat{p} = \frac{\text{successes}}{\text{total}}$
⇒ draw a picture 

③ Calculate test statistic

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \text{ or } \frac{\bar{x} - \mu}{SE} \quad \text{where } SE = \frac{s}{\sqrt{n}}$$

④ Make decision & interpret
* if p-val $> \alpha$ accept H_0
* if p-val $< \alpha$ reject H_0

(do check if $n \geq 30$ bcz if not use t-dist.)

~~* Type 1 error is more costly~~

Lecture 9 and 10

* Difference b/w two sample proportions:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

* Confidence interval: $CI = (\hat{p}_1 - \hat{p}_2) \pm z \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

*

* Pooled estimate of proportion: $\hat{p} = \frac{(\text{no. of success}) + (\text{no. of succ}}{n_1 + n_2}$

* z-value: $\frac{(\hat{p}_1 - \hat{p}_2)}{SE}$ where $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

* comparing two proportions:

population parameter : $p_1 - p_2$

point estimate : $\hat{p}_1 - \hat{p}_2$

* for CI (confidence interval) use: $\hat{p}_1$ & $\hat{p}_2$

* for HT (hypothesis testing) use:

when $H_0: p_1 = p_2$, use $\hat{p}_{pool} = \frac{\#suc_1 + \#suc_2}{n_1 + n_2}$

when $H_0: p_1 - p_2 = (\text{some value besides } 0)$, use $\hat{p}_1$ and $\hat{p}_2$

* mean standard error for two samples: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Lecture 11 and 12

* Chi-Square test for Goodness of fit: (χ^2)

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$ where k = total no. of cells.

* degree of freedom (df): $df = k - 1$

* Chi-Squar test of Independence: (hypothesis test)

(2 $\geq$ variables) hypothesis: H_0: Attributes are independent

H_A: " are associated

* Test statistic:

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where $df = (R-1)(C-1)$ (R: #rows, C: #cols)

* Expected values = $\frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$

Lecture 13 and 14

* t-distribution : $t = \frac{\bar{x} - \mu}{SE}$, $SE = \frac{s}{\sqrt{n}}$

* when to use z-test and when t-test?
  → If standard deviation of population is unknown, z-test should be used.
  → If standard deviation of population is known, then the size of sample is used to determine : → if $n < 30$, t-test is used.
→ In z-test you know population SD but in t-test you know sample SD.

* Paired t-test : ① first we draw out a new column called "difference" from given two columns that need to be used.
② Find $\bar{x}_{diff} = \frac{\Sigma \bar{x}_{diff}}{n}$, $SE = \frac{s_{diff}}{\sqrt{n}}$ where

$s_{diff} = \sqrt{\frac{\Sigma (x_d - \bar{x}_d)^2}{n_d}}$, $t = \frac{\bar{x}_{diff}}{SE}$, $df = n - 1$

③ find p value & determine results.

Lecture 15

* Difference in two means (t-test)

point estimate : $\bar{x}_1 - \bar{x}_2$

null value : $\mu_1 - \mu_2$

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = (n_1 - 1, n_2 - 1) \leftarrow \text{minimum from both is selected}$$

* ANOVA (Analysis of Variance)

→ hypothesis : H_0: All means are equal

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \ldots$

~~H_A : All means are not equal~~

~~$H_A : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \ldots$~~

H_A : Atleast 1 mean is different.

→ test statistic :

$$F = \frac{\text{variability b/w groups}}{\text{variability within groups}}$$

ANOVA table:

	df	sum squared (SS)	mean squared (MS)	F-score	p-value
(G) group (b/w)	$K-1$	$SSG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2$	$MSG = \frac{SSG}{df_G}$	$F = \frac{MSG}{MSE}$	
(E) error (within)	$df_T - df_G$	$SSE = SST - SSG$	$MSE = \frac{SSE}{df_E}$		
(T) total	$n-1$	$SST = \sum_{i=1}^{K} (x_i - \bar{x})^2$			

Lecture 16

* Multiple comparisons : testing many pairs of groups
* Bonferroni correction : $\alpha^* = \alpha / K$ where K is number of comparisons being considered.
$K = \frac{K(K-1)}{2}$ (all possible pairs of groups)
* One-way ANOVA test :
* $s_{pooled} = \sqrt{MSE}$
* $SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$, $Df_E = (\bar{x}_1 - \bar{x}_2)$ / SE

Lecture 17

* Residuals (e) : $e_i = y_i - \hat{y}_i$ (actual, predicted)
* Least squares line : $\hat{y} = \beta_0 + \beta_1 x$ where
$\beta_1 = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$, $\beta_0 = \bar{y} - b\bar{x}$
* conditions for least squares line :
① linearity ② nearly normal residuals
③ constant variability ④ No extreme outliers

Lecture 18

* Coefficient of corelation : ~~$r = \frac{\Sigma[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2}}$~~

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

* range of r is b/w -1 and 1

* Types of outliers in linear regression:

① High Leverage (horizontally away from blob)

② Influential (influce the slope of regression)

Lecture 19, 20, 21

* multi-linear regression line : $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

for two variables : $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 ,$$

$$\beta_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$\beta_2 = \frac{(\Sigma x_2^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

where only used to scale down the huge values

$$\Sigma x_1 y = \Sigma x_1 y - \frac{(\Sigma x_1)(\Sigma y)}{N}$$

$$\Sigma x_2 y = \Sigma x_2 y - \frac{(\Sigma x_2)(\Sigma y)}{N}$$

$$\Sigma x_1 x_2 = \Sigma x_1 x_2 - \frac{(\Sigma x_1)(\Sigma x_2)}{N}$$

Lecture 21, 23, 22

* Multi-colinearity: types: Structural, Data
  we build these from data ↙ We observe ↘ these from data
* VIF (variance Inflation Factors)
* VIF ranges from 1 to $+\infty$
* Interpretation: → VIF : 1 indicates no correlation
→ VIF b/w 1 and 5 : indicates moderate correlation
→ VIF greater than 5 indicate severe corelation
* Formula for each coefficient of the least square regression line is calculated like this: $VIF_i = \frac{1}{1 - R_i^2}$ where i indicates the number of ~~vari~~ coefficient i.e $\beta_0, \beta_1, \beta_2, \beta_3, \ldots$ etc.
where $R^2 = (R)^2 = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$

Lecture 24

* Adjusted R^2 : $R^2_{adj} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$

where N is total sample size and p is the number of independent variables.

* Model Selection methods: * Forward-Selection
* Backward-elimination.

Lecture 25

* Modelling condition: ① residuals are nearly normal
② residuals have constant variability
③ residuals are independent (no patterns appear)
④ each variable is linearly related to the outcome
* Options to improve model: ① Transforming variables: Log transformation, square root, inverse, truncation.
② adding new variables ③ use more advanced methods.
* Logistic Regression: SLR, MLR (covered)
* Odds: $\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1-P(E)}$
* Logit function: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ for $0 \leq p \leq 1$

Lecture 26

* Confidence Interval using PE (point estimates) & SE
→ Log odds ratio: $CI = PE \pm \underbrace{CV}_{\text{z-value}} \times SE$
→ odds ratio: $\exp(CI) = \exp(PE \pm CV \times SE)$.

Lecture 27 + 28

* Sensitivity, Specificity

Predicted \ Actual		
	True +ve	False +ve
	False -ve	True -ve

$\text{Sensitivity} = \frac{TP}{(TP+FN)}$, $\text{Specificity} = \frac{TN}{(TN+FP)}$

$\text{positive predicted value} = \frac{TP}{(TP+FP)}$, $\text{negative predicted value} = \frac{TN}{(FN+TN)}$