ST1131 Summary 24/25

Teh Xu An

Variables

Types of Variables

Types	Category	Remarks	Examples
Nominal	Categorical	No Order	Race, Gender
Ordinal	Categorical	Ordered	Ratings
Discrete	Quantitative	Numbers with fixed difference	Number of ..., Set of integers
Continuous	Quantitative	Forms intervals	Age, Height, Weight

Quantitative → Categorical Ordinal: Ordered ranges of values

Variable Roles (Response / Explanatory)

Response	Alt (LRM): Dependent, Target, Output
Explanatory	Alt (LRM): Regressor, Independent, Predictor, Covariate

If Explanatory is Categorical, groups are compared.
If unable to identify roles of variables, explore association.

Lurking / Confounding Variables

Lurking	Unobserved & influential, potential confounding
Confounding	2 Observed but related variables; undifferentiable

Basis of Lurking /Confounding: Correlation does not imply causation

Relative Frequencies

Proportion (of X)	P(X)=Observations(of X) / Total Observations
Percentage (of X)	Proportion (of X) × 100%

Studies & Sample Survey

Study Types (Observational / Experimental)

Observational	Observe subjects for variables, no interventions Easier to conduct, cost effective
Experimental	Assign subjects to treatments, record observations Able to control for lurking variables May be unethical, impractical, or costly

Steps of a Sample Survey

Description	A study that records sample's subjects' response/ measurements, drawn from the population randomly
Step 1	Identify the Population
Step 2	Create Sampling Frame, a list of subjects in population to sample, ideally the whole population
Step 3	Set Sampling Design for selecting subjects from the sampling frame
Step 4	Collect Data from the randomly selected subjects

Random Sampling

Premise	Good sampling designs employ randomization ie chance over convenience
Description	Each sample of size *n* has the same chance of being selected from a sampling frame
Step 1	Subjects in sampling frame are numbered
Step 2	Generate a set of *n* random numbers
Step 3	Subjects with numbers in the set of *n* numbers are picked to be a Simple Random Sample

Other Random Samples: Clustered / Stratified

Biases in Sample Surveys

Sampling Bias	Caused by sampling frame/design step
Non-Sampling Bias	Response / Non-Response Bias.

Sampling Bias: Under Coverage, Non-Random Sample
Response Bias: Subject wrong response, Misleading questions
Non-Response Bias: Cannot be reached, Refuse to participate

Large Sample Size does not guarantee an unbiased sample!

Poor alternative Surveys to Sample Surveys

Convenience	Sample selected based on ease of access
Volunteer	Subjects are encouraged to participate

Elements of good Experimental Studies

Control	A group without treatment for comparison
Randomization	Random assignment of treatment
Blind	Subjects are unaware of treatment or placebo
Double Blind*	Both Subjects and Administrators are unaware of treatment or placebo.

Role of Randomization in Experimental Studies

Eliminate Bias that may appear if we assign subjects by hand

Balance groups on lurking variables that we know affects response / that may be unknown to us

Numerical Summaries

Center	Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, sensitive to outlier, use when not skewed, $Y = bX + a \to \bar{Y} = b\bar{X} + a$ Median: 50th quantile, robust to outlier, use when skewed
Variability	Range: (min, max), always used Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, used when symmetric Standard Deviation: $s = \sqrt{s^2}$, used when not symmetric
Variability (Notes)	$Y = bX + a \to s_Y^2 = b^2 s_X^2, s_Y = \lvert b \rvert s_X$ $\sim 68\%$ in $\bar{X} \pm s$, $\sim 95\%$ in $\bar{X} \pm 2s$, $\sim 99.7\%$ in $\bar{X} \pm 3s$ Works only in unimodal symmetric models.
IQR	Range of $(q_{0.25}, q_{0.75})$ or (Q_1, Q_3)
Quantile	q_p: $100p$-th quantile ie $100p$ percent fall below q_p
Quartile	Q_1, Q_2, Q_3: $q_{0.25}, q_{0.5}, q_{0.75}$ / lower, median, upper
Five-Number Summary	Includes $q_0, q_{0.25}, q_{0.5}, q_{0.75}, q_1$ Good indicator of center and variability
Outliers	$X_{outlier}: < Q_1 - 1.5 \times IQR$ or $> Q_3 + 1.5 \times IQR$

Probability Topic

Sample Space	Set of all possible outcomes
Event	Subset of sample space

Suppose events A, B and C are in sample space S

$$P(A) = \frac{\text{number of outcomes in A}}{\text{total number of possible outcomes in S}}$$

$P(A) \geq 0$ | $P(S) = 1$ | $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

A & B are independent $\equiv \big(P(A \cap B) = P(A)P(B)\big) \vee \big(P(A|B) = P(A)\big)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B|A)}{P(B)}, \text{ where } P(B) > 0$$

Sensitivity	$P(+\|D)$	Specificity	$P(-\|D^c)$
Prevalence	$P(D)$		

$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$, where $B_1, B_2, \dots, B_n$ partitions S

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}, \text{ where } B_1, B_2, \dots, B_n \text{ partitions S}$$

Random Variables

Definition	Measurement of the outcome of an experiment
Probability Distribution	Specifies possible values of a random variable and their probabilities

Discrete Random Variables

Definition	Takes on a set of separate values
Probability Distribution	Assigns a probability p_X to each possible values of X
Annotation notes	Uppercase letters: Denotes the random variable Lowercase letters: Denotes the value it takes on
Mean	$\mu = \sum_x x p_x, E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$
Variance	$\sigma^2 = \sum_x (x - \mu)^2 p_x, Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$

Continuous Random Variables

Mean	$\mu = \int x f(x) dx$	Quantile	$P(X \leq q_p) = p$
Variance	$\sigma^2 = \int (x - \mu)^2 f(x) dx$		

Binomial Distribution $\big(X \sim Bin(n, p)\big)$

Definition	n independent trials with p probability of success		
Bernoulli	$Bin(1, p)$, a binomial distribution with only 1 trial		
P(x successes)	$P(X = x) = C_x^n p^x (1 - p)^{n-x}$		
Mean of X	$E(X) = np$	Variance	$Var(X) = np(1 - p)$

Normal Distribution (Gaussian Distribution / $X \sim N(\mu, \sigma^2)$)

Definition	Symmetric, bell-shaped / unimodal, and characterized by its mean μ and its variance σ^2
If $d > 0, P(X \leq \mu - d) = P(X \geq \mu + d)$	$q_{1-p} = 2\mu - q_p$

Suppose $X \sim N(\mu_X, {\sigma_X}^2)$ and $Y \sim N(\mu_Y, {\sigma_Y}^2)$,

General Linear Transform	$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$
Standardize Normal	$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$, where $X \sim N(\mu, \sigma^2)$
Binomial Estimation	$Bin(\text{moderately large } n, p \approx 0.5 \pm 0.2) \approx N\big(np, np(1 - p)\big)$, where $np(1 - p) \geq 5$

t-Distribution with $df = n - 1$

t-score	$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$; @ $n \geq 30$, $t_{n-1} \approx z$, $\frac{1}{\sqrt{n}} t_{n-1,1-\alpha/2} \approx q_{1-\alpha/2}$

Sampling Distribution

Quantitative Sampling Distribution	Sample Mean, $\bar{X} \to \mu$
$Bin(1, p)$ Sampling Distribution	Sample Proportion, $\hat{p} \to p$

Central Limit Theorem

Suppose IID $X_1, X_2, \dots, X_n$, and $n \geq 30$, then $\bar{X} \sim N(\mu, \sigma^2/n)$

Confidence Interval

Point Estimate	Single best guess number for population parameter Sample proportion: $\hat{p}$, Sample mean: $\bar{X}$
Interval Estimate	An interval of numbers within which the parameter value is believed to fall

ST1131 Summary 24/25

Teh Xu An

Estimates	$\bar{X} \approx \mu$	$s^2 \approx \sigma^2$	$s \approx \sigma$	$X_{(0.5)} \approx q_{0.5}$
CI level α	$\hat{p} \pm q_{1-\alpha/2} \times \sigma$ or $\bar{X} \pm t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$			
Interpret CI	$(1-\alpha) \times 100\%$ confident that $\hat{p}/\bar{X}$ falls in CI level α			
CI Width	$2 \times q_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or $2 \times t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$			

Hypothesis Test	
Assumptions To Check	Random sample data, Normal Distribution Stop if assumptions not fulfilled
Test Statistic	Score of the sample against H_0
Null Distribution	Distribution of test statistic under H_0
p-value	probability of sample against null distribution
Type I Error	Reject H_0 when it is true AKA false positive
Type II Error	Don't reject H_0 when it is false AKA false negative
Test Power	$1-\beta$, where β is the probability of Type II Error
Independence & Dependence	
Independent Sample	Observations in one sample implies nothing in another sample
Dependent Sample	Two groups/samples comprise the same set of subject/individuals, hence related samples
Independent Samples, Equal Variance (Two sample t-test)	
Assumptions	+ Hypothesis Assumptions Both population variances are the same, $\frac{s_X^2}{s_Y^2} \in \left(\frac{1}{3}, 3\right)$
Variance Test	$H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0 / \mu_x - \mu_y < 0 / \mu_x - \mu_y > 0$
Supposition	Let $X \sim N(\bar{X}, s_X^2)$ of size n_1, $Y \sim N(\bar{Y}, s_Y^2)$ of size n_Y Since both population equal variance, let σ^2 be that.
Pooled Variance	$\sigma^2 = s_p^2 = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}$
Test Statistic	$T = \frac{(\bar{X}-\bar{Y})-0}{se}$ where $se = s_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$ Under H_0, T follows t-distribution $df = (n_X+n_Y-2)$
p-value	$H_1: \mu_x - \mu_y \neq 0$, Two tail probability from $t_{n_X+n_Y-2}$ $H_1: \mu_x - \mu_y > 0$, Right area of T from $t_{n_X+n_Y-2}$ $H_1: \mu_x - \mu_y < 0$, Left area of T from $t_{n_X+n_Y-2}$
Conclusion	Interpret p-value
Independent Samples, Unequal Variance (Welch Test)	
Assumptions	Two sample t-test minus same variance
Test Statistic	Test Statistic same as Two Sample t-test Under H_0, T follows t-distribution df (too complex)
df Calculation (Bonus)	$df = \frac{(se_X+se_Y)^2}{\frac{se_X^2}{n_X-1}+\frac{se_Y^2}{n_Y-1}}, se_X = \frac{s_X^2}{n_X}, se_Y = \frac{s_Y^2}{n_Y}$
Dependent Samples (Dependent t-test for paired samples)	
Premise	Every observation in a sample has a matched value in other sample, hence we compare the mean of differences of matched observations with 0, then we use one sample t-test

Linear Regression	
Linear Regression	$Y = \beta_0 + \beta_1 X + \epsilon$, ϵ is a random variable with variance σ^2, β_0 is Y-intercept, β_1 is slope of line $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$, where σ^2 is constant
Indicator Terms	$I(X = value) = \begin{cases} 1, & X = value \\ 0, & X \neq value \end{cases}$
Interaction Terms	If $Cor(X_1, X_2)$ is high or X_1 and X_2 are associated, we can have interaction term $(X_1 * X_2)$ as a regressor
"Linear"	Linearity in the parameters
OLS Estimation	Ordinary Least Square Estimation Minimises sum of squared residuals, e_i's: $e_i = Y_i - \hat{Y}_i$
t test	Test significance of one regressor (or one coefficient)
F test	Test significance of the whole model
t-test for β	
Step 1	Check assumptions
Step 2	$H_0: \beta = 0$ or H_0:regressor X is no significant $H_1: \beta \neq 0$ or H_1:regressor X is significant *One-sided tests are also possible
Step 3	$t = \hat{\beta}/SE(\hat{\beta})$, null distribution of t is t_{n-2}
Step 4	Derive p-value found from R output
Step 5	Conclude whether the slope β is significantly different from 0 at a pre-specified α-level
F-tests in a Linear Model	
Hypothesises	H_0: all the coefficients, except intercept, are zero H_1: >0 coefficients, except intercept, are nonzero
Fixes	Linear assumption violated: Transform Regressor Variance not constant: Transform Target
Standardized Residual	$SR = \frac{Y-\hat{Y}}{SE(Y-\hat{Y})}$, where $SE(Y-\hat{Y}) = \frac{\sigma_{Y-\hat{Y}}}{\sqrt{n}}$
Checks for Linear Model Assumptions	
Assumptions of LRM $Y \sim X$	Random Data, Relationship between X and Y is linear Error term $\epsilon \sim N(0, \sigma^2)$, homoscedasticity = constant σ
Constant σ^2	Histogram and QQ plot of SR are normal
Normality	Plot of SR against $\hat{Y}$ and SR versus X: points scatter randomly about 0, within the interval (-3, 3)
Outlier	Outlier Points: $\|SR\| > 3$
Influential	Influential Points: $cook.distance > 1$
Linear Model Interpretations	
Coefficient of determination R^2	Proportion of total variation of the response that is explained by the model, falls between 0 and 1. Big R^2 = good fit, but more variables will increase R^2
Adjusted R^2 R^2_{adj}	$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$ Where k is the number of regressors in the model
$Cor(X,Y)$&R^2	$\sqrt{R^2} = \|Cor(X,Y)\|$

Tables	
Frequency Tables	
Summary (To Include)	Modal Category Relative Frequency
Contingency Table	
Summary (To Include)	Modal Category Relative Frequency

Plots	
Bar Plots	
Summary (To include)	+ Frequency Tables Summary Mention groups of high/low proportions Any trends with ordinal categories.
Clustered / Stacked for comparing 2 categorical variables.	
Histogram	
Summary (To include)	Clusters, Gaps, Deviations, Suspected Outliers Mounds: Unimodal, Bimodal, Multimodal Skew: Symmetric, Left-Skew, Right-Skew
Remarks (Skews)	Longer = Thicker = Heavier Shorter = Thinner = Lighter Histogram is skewed towards the longer tail.
Boxplot	
Summary (To include)	Box: Lower Bound, Line, Upper Bound: Q_1, Q_2, Q_3 Whiskers: Lower Whisker, Upper Whisker: $Q_1 - 1.5 \times sd, Q_3 + 1.5 \times sd$ Outliers: Points beyond the whiskers
Remarks	Useful for identifying potential outliers Indicator of skewness if unimodal
Scatterplot	
Summary (To include)	Correlation: $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$
Remarks	Useful for identifying potential outliers Indicator of skewness if unimodal
QQ Plot	
Description	To check normality by plotting standardized sample quantiles against theoretical quantiles of a $N(0,1)$
Summary (To Include)	Tail above/below Line: Shorter/Longer than Normal Normality
Remarks	If the points follow a straight line, there is evidence that the data came from a normal distribution