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	Variable on which comparisons are made Alt (LRM): Dependent, Target, Output		
Explanatory	Any variable the response might depend on Alt (LRM): Regressor, Independent, Predictor, Input, 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	Observed but correlated with another explanatory variable ie unable to determine which variable is causing a change in response; Undifferentiable		
Basis of Lurking /Confounding Variables: Correlation does not imply causation			

Relative Frequencies	
Proportion (of X)	Observations(of X) / Total Observations
Percentage (of X)	Proportion (of X) × 100%

Studies & Sample Survey		
Study Types (Observational / Experimental)		
Observational	The variables are observed for sampled subjects, without anything done to them	
	Easier to conduct	
Experimental	Conducted by assigning subjects to certain experimental treatments and then observing the outcome on the response variable	
	Able to control for lurking variables May be unethical, impractical, or costly	
Steps of a Sample Survey		
Description	A study that asks questions/take measurements of the subjects in a sample drawn from the population randomly	
Step 1	Identify the Population	
Step 2	Compile a list of subjects in the population from which the sample will be taken ie Sampling Frame, ideally lists all subjects in population	
Step 3	Specify a method for selecting subjects from the sampling frame ie Sampling Design	
Step 4	Collect Data	

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	A result of sampling design step, or sampling frame step
Non-Sampling Bias	Response / Non-Response Bias. Not a result of sampling design.
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: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Median: 50th quantile
Center (Notes)	$Y = bX + a \rightarrow \bar{Y} = b\bar{X} + a$ Outliers: Mean is sensitive, Median is robust Skewed: Median, Not Skewed: Mean
Variability	Range: [min, max] Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ Standard Deviation: $s = \sqrt{s^2}$
Variability (Notes)	$Y = bX + a \rightarrow s_Y^2 = b^2 s_X^2, s_Y = \lvert b \rvert s_X$ ~68% in $\bar{X} \pm s$, ~95% in $\bar{X} \pm 2s$, ~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
Use Cases	Range: Always Variance and sd: Approximately bell-shaped
	IQR: Not bell-shaped
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		
Event (A & B) Combination	Union ($A \cap B$): Belong to either or both Events Intersection ($A \cup B$): Belong to both Events		
Complement	Exclusive of Event, Subset of all not in Event		
Probability	Proportion of times an Event occurs		
Suppose events A, B and C are in sample space S			
$P(A) = \frac{\text{number of outcomes in A}}{\text{total number of possible outcomes in S}}$			
$P(A) \geq 0$	$P(S) = 1$	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$			
A & B are independent $\equiv \big(P(A \cap B) = P(A)P(B)\big) \vee \big(P(A \mid B) = P(A)\big)$			
$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, where $P(B) > 0$		$P(B \mid A) = \frac{P(B) \times P(A \mid B)}{P(A)}$	
Sensitivity	$P(+ \mid D)$	Specificity	$P(- \mid D^c)$
Prevalence	$P(D)$		
$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$, where $B_1, B_2, \ldots, B_n$ partitions S			
$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$, where $B_1, B_2, \ldots, B_n$ 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 $(X \sim Bin(n, p))$			
Definition	*n* trials, either success or failure Each trial has the same probability of success *p* The *n* trials are independent		
Bernoulli	$Bin(1, p)$, a binomial distribution with only 1 trial		
Probability of *x* successes	$P(X = x) = C_x^n p^x (1 - p)^{n-x}$		

ST1131 Summary 24/25

Teh Xu An

Mean of X	$E(X) = np$	Variance	$Var(X) = np(1-p)$

Poisson Distribution with parameter λ	
Definition	Follows $P(X=k) = \frac{e^{-\mu}\mu^k}{k!}, k = 0,1,2,\ldots$ Where e is approximately 2.71828, λ is the expected no. events per time unit and $\mu = \lambda t$ is the number of events over time period t.
Binomial Estimation	A Binomial distribution with large n and small p can be accurately approximated by a Poisson distribution with parameter $\mu = np$
	$Bin(\text{large } n, \text{small } p) \approx Pois(np/t)$
Mean of X	$\mu = np$
Variance	$Var(X) = np(1-p) \approx np$

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 $d > 0, P(X \le \mu - d) = P(X \ge \mu + d)$	$q_{1-p} = 2\mu - q_p$
Suppose $X \sim N(\mu_X, {\sigma_X}^2)$ and $Y \sim N(\mu_Y, {\sigma_Y}^2)$,	
Add Constant	$X + a \sim N(a + \mu_X, \sigma_X^2)$
Add Normal	$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
Multiply Normal Variable	$X_1 + X_2 + \cdots + X_n = \sim N(n\mu, n\sigma^2)$ Where $X_1, X_2, \ldots, X_n$ are independently identically distributed (IID) $N(\mu, \sigma^2)$.
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 $X \sim N(\mu, \sigma^2)$ Then Z is the Z-score of X
Binomial Estimation	The Binomial distribution with a moderately large n and p not close to 0 or 1, then $Bin(n,p)$ tends to be symmetric and well approximated by normal distribution $N\big(np, np(1-p)\big)$, where $np(1-p) \ge 5$
	$Bin(\text{moderately large } n, p \approx 0.5 \pm 0.2) \approx N\big(np, np(1-p)\big)$, where $np(1-p) \ge 5$
t-Distribution with $df = n - 1$	
t-score	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$; @ $n \ge 30, t_{n-1} \approx z, \frac{1}{\sqrt{n}} t_{n-1,1-\alpha/2} \approx q_{1-\alpha/2}$

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 $X_1, X_2, \ldots, X_n$, and $n \ge 30$, 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			
Estimates	$\bar{X} \approx \mu$	$s^2 \approx \sigma^2$	$s \approx \sigma$	$X_{(0.5)} \approx q_{0.5}$
General Form of CI	point estimate $\pm$ margin of error Where margin of error measure how accurate the point estimate is likely to be in estimating a parameter			

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	
Hypothesis Definition	A statement about a population, usually claiming that a parameter takes a particular numerical value or falls in a certain range of values.
Significance Level	Denoted as α A number such that we reject H_0 when p-value $\le \alpha$
Step 1	Check Assumptions: Random sample data, Normal Distribution Stop if assumptions not fulfilled
Step 2	Null Hypothesis: parameter takes a value ie $H_0: v = v_0$ Alternative Hypothesis: parameter falls in some other range of values ie $H_1: v \ne v_0 \,/\, v < v_0 \,/\, v > v_0$
Step 3	Test statistic: score of the sample against H_0 Null Distribution: Distribution of test statistic under H_0
Step 4	p-value: probability of sample against null distribution Small p-value is strong evidence against H_0
Step 5	If p-value $\le \alpha$, reject H_0, otherwise, do not reject H_0
Type I Error	Reject H_0 when it is true AKA false positive
Type II Error	Do not 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	Quantitative response variable for both groups Two samples are independent Population distribution of each group is approximately normal, especially for $n \le 30$ Both population variances are the same
Variance Test	Null Hypothesis: Two samples are from two populations with the same variance ie $H_0: \mu_x - \mu_y = 0$ Alternative Hypothesis: ie $H_1: \mu_x - \mu_y \ne 0 / \mu_x - \mu_y < 0 / \mu_x - \mu_y > 0$
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 $df = (n_X + n_Y - 1)$
p-value	$H_1: \mu_x - \mu_y \ne 0$, 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	Same as Independent Samples, Equal Variance, except test of equal variance is significant ie reject $H_0: \mu_x - \mu_y = 0$
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
df Calculation	$df = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{s_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{s_Y^2}{n_Y}\right)^2}{n_Y - 1}}$
	$df = \frac{(n_X-1)(n_Y-1)}{(n_X-1)c_Y^2 + (n_Y-1)c_X^2}$, where $c_X = \frac{\frac{s_X^2}{n_X}}{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}, c_Y = 1 - c_X$
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	
Regression	A regression of response Y on the regressor X Mathematical relationship between the mean of Y and different values of X.
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	$I(X = value) = \begin{cases} 1, & X = value \\ 0, & X \ne value \end{cases}$
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 $H_1: \beta \ne 0$ 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: at least one of the coefficients, except intercept, are nonzero

Fixes	Linear assumption violated: Transform Regressor Variance not constant: Transform Target
Standardized Residual	$SR = \frac{Y - \hat{Y}}{SE(Y - \hat{Y})}$
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: $\lvert SR \rvert > 3$
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. Bigger R^2 means better goodness of fit of the model, however, 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
$Cor(X,Y) \& R^2$	$\sqrt{R^2} = \lvert Cor(X,Y) \rvert$

Tables	
Frequency Tables	
Data Types	Column: Categorical Data Entries: Quantitative
Summary (To Include)	Modal Category Relative Frequency
Contingency Table	
Data Types	Column / Row: Categorical Entries: Quantitative
Summary (To Include)	Modal Category Relative Frequency

Plots	
Bar Plots	
Description	Display vertical bars for each category, with height proportional to their frequencies.
Data Types	X-axis: Categorical Y-axis: Continuous Quantitative
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	
Description	Uses bars to portray frequencies of possible outcomes (equal range) of a quantitative variable.
Data Types	X-axis: Quantitative → Ordinal Categorical Y-axis: Continuous Quantitative
Summary	Clusters, Gaps, Deviations, Suspected Outliers

(To include)	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	
Description	Visual Representation of Five-Numbers Summary
Data Types	X-axis: Categorical Y-axis: Quantitative
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	
Description	Comparison of 2 quantitative variables.
Data Types	X-axis & Y-axis: Quantitative
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 $N(0,1)$
Data Types	X-axis & Y-axis: Quantitative
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

Useful R Codes		
Import Data	read.csv	..., show_col_types=, sep=, header=
	data.frame	..., row.names=, check.rows=, check.names=, fix.empty.names=, stringAsFactors=
Data Write/ Data Read	length	..., <- value
	ifelse	test, yes, no
	names	..., <- value
	rownames	..., do.NULL=, prefix=, <- value
	columnnames	..., do.NULL=, prefix=, <- value
	df[x,y]/df$x	Filter list(names)/number, can -
	df$out	Get outliers
	df$group	Get outliers in each group
	df$names	Get names
	which	pred(df)
	cor	Get correlation
	summary	object, maxsum=, digits=, quantile.type=, ...=
	apply	X, margin, FUN, ..., simplify=
Tables	table	..., exclude=, useNA=, dnn=, deparse.level=, x=, row.names=, responseName=
	prop.table	..., margin=
Plots	plot	..., type=p/l/b/c/o/h/s/S/n, main=, sub=, xlab=, ylab=, asp=, col=, legend.text=
	barplot	beside=, legend.text=, density=, col=, main=, xlab=, ylab=, xlim=, ylim=, xpd=, formula=, data=, subset=, na.action=
	pie	..., labels=, clockwise=, init.angle=, angle=, col=, main=
	hist	..., breaks=, freq=, probability=, include.lowest=, right=, col=, main=, xlab=, ylab=, xlim=, ylim=, axes=, labels=
	boxplot.stats	x, coef=, do.conf=, do.out=
	boxplot	x, data=, subset, xlab=, ylab=, ann=, col=
	abline	a=, b=, h=, v=, reg=, coef=, untf=
	qqnorm	..., ylim=,
	qqline	..., datax=, distribution=
Evaluate Distribution	rbinom	n, prob, size=
	dbinom	x, prob, size=, log=
	pbinom	q, prob, size=, lower.tail=, log.p=
	qbinom	p, prob, size, lower.tail=, log.p=
	rpois	n, lambda
	dpois	x, lambda, log=
	ppois	q, lambda, lower.tail=, log.p=
	qpois	p, lambda, lower.tail=, log.p=
	rnorm	n, mean=, sd=
	dnorm	x, mean=, sd=, log=
	pnorm	q, mean=, sd=, lower.tail=, log.p=

	qnorm	p, mean=, sd=, lower.tail=, log.p=
	sample	x, size, replace=, prob=
	replicate	n, expr
Tests	t.test	x, y=, alternative=, mu=, paired=, var.equal=, conf.level=
	var.test	x, y, ratio=, alternative=, conf.level=
LRM	lm	Formula, data, subset=, weights=, na.action=, method=, x=, y=, qr=, ...
	summary.lm	...
	predict.lm	...
	lm$res	Get list of residuals
	rstandard	...
	cook.distance	...
	predict	..., newdata=, na.action=