Ling – ST1131 cheatsheet

	Population statistics	**Sample statistics**
Total Number	N	n
Mean	μ (given)	$\bar{x} = \frac{\Sigma x}{N}$
Median	-	$\tilde{x}$ (middle)
Standard deviation	$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or given	$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$
Pearson Coefficient of linear correlation	ρ: $\sigma^2{}_{x\pm y} = \sigma^2{}_x + \sigma^2{}_y \pm 2\rho\sigma_x\sigma_y$	$r = \frac{1}{n-1}\Sigma(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_x})$
Proportion	p (given)	$\hat{p} = \frac{X}{n}$
Calculated test statistic	-	z, t, $\hat{p}$, Binomial

Rules	Formula	Remark
68-95-99.7 rule	$\mu \pm 1{,}2{,}3\ \sigma$	Can be used to find extreme data
Interquartile Range	IQR = Q3 - Q1	Can be used to test spread
Central location	Median	Compare between two datasets
5 Numbers Box Plot	Min|Q1|Median|Q3|Max	Spread & Outliers
Outliers	>Q3 + 1.5IQR <Q1 - 1.5IQR	If outliers' are present, distribution is less likely to be normal
Influential Point	Recalculation of r value	Is an outlier and removing it affect the best fit line
Simpson's Paradox	An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. Due to the difference in weightage	
Restricted Range	Association changes when data range changes	
Average data	Correlation based on average data most likely resulting in higher association	
Central Limit Theorem	For sample with large sample size, the sampling distribution of the sample mean is approximately a normal distribution	Note it is the sample mean, not the sample
Law of Large Numbers	With a large number of experiments, the average will tends towards the expected value	
Binomial probability	$P(x) = \frac{n!}{x!(n-x)!}p^x 1 - p^{n-x}$	$\frac{n!}{x!(n-x)!} = nCx$
Normality Test	Anderson Darling Normality Normality probability plot	P≥0.05 when normal Linear plot
Examine graph	Overall pattern (Form/ Direction/ Strength) Deviations (Outliers)	
Confidence Interval	α% of the confidence intervals constructed in this way would contain the true value for the population parameter of H_0/H_1. We estimate with α% confidence that between CI of H_0/H_1	
Conclusion statement	There is sufficient/insufficient evidence to support H_0/H_1	

Studies	Method	Remarks
Observational Studies	Observe samples without modification	Contrast to experimental studies Placebo/Control Group/Blinding
Simple Random Sampling	Every sample has equal probability	Essential for N, Z, T tests
Stratified Random Sampling	Divide sample into representative groups before SRS	Useful if the sample contains different groups
Cluster Random Sampling	Divide sample into groups then randomly select a few groups	Low cost and no need sampling frame Large sample size required to reduce margin of error
Voluntary response Sampling	Require sample to voluntarily response	Subject to response/ non-response bias
Multistage Sampling	Conduct the sampling in different stages	Easily mistaken with stratified sampling

Ling – ST1131 cheatsheet

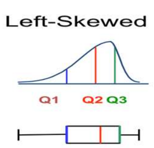
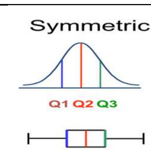
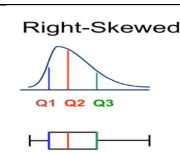
CI (α/2) sample mean	90%	95%	99%
CI = ± Z* × $\frac{\sigma}{\sqrt{n}}$	1.645	1.960	2.576

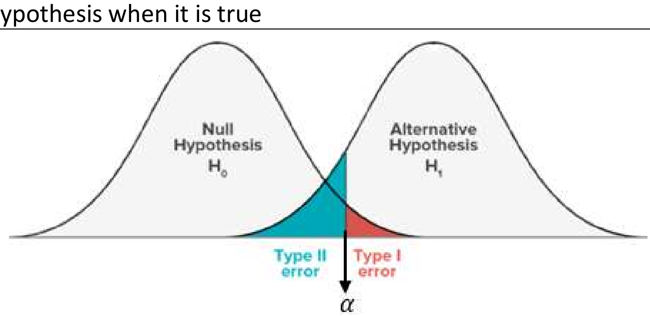
Biasness	**Problem**		
Response Bias	Misleading questions, incorrect response		
Non-Response Bias	Subjects cannot be reached or refused to participate		
Sampling Bias	Under coverage/ Non-random sample		
Anecdotal data	Anecdotal evidence is based on haphazardly selected individual cases		
Bias	Affects Mean	**Variability**	Affects Spread

Linear Transformation	**Center, $\bar{x}$**	**Spread, s or σ**
Addition (a)	a $\bar{x}$	no change
Multiplication (b)	b $\bar{x}$	$b^2 s^2$

	Assumption	**Implication**
Z, T test	Normality	All formulas require the sample to be normally distributed
Z, T test	Simple Random Sample (Independence)	Result can only apply to selected samples and not the population

Interpret Association	**Quantitative**	**Categorical variable**
	Scatter diagram	Two-Way Table

			
Outliers	Left Mean < Median	No Mean = Median	Right Mean > Median

Type I Error	**Type II Error**
P(Reject H_0| H_0 is true) = significant level (α) α times there will be an error	P(Accept H_0| H_a is true) = β Power = 1 - β = P(Reject H_0| H_a is true), Higher power = better
A Type I error occurs when the researcher rejects a null hypothesis when it is true	A Type II error occurs when the researcher accepts a null hypothesis that is false
	Both errors are calculated using **Z test** Despite their initial distributions The probability of *not* committing a Type II error is called the Power of the test. Power helps to determine if sample size is large enough. If your sample size is too small, your results may be inconclusive when they may have been conclusive with a large enough sample. It's better to commit Type II error than Type I error

Regression	**Implication**	**Remark**
R-Squared	Fraction of the variation in the values of y that is explained by the least-squares regressions of y on x	
Least-squares	Least square is when the minimum sum of residual^2	
Correlation, r	Strength and direction of the <u>linear</u> association between two quantitative variables Non linear r/s: Circular r/s, Curve r/s	Independent of response/explanatory variable Not resistant to outliers (e.g. influential observation) Independent of unit of variables
Residual	Sum of residual = 0, above 0 = overestimated, under 0 = underestimated	

Ling – ST1131 cheatsheet

	Residual = $y - \hat{y}$, vertical distance between actual and predicted response variable Regression Line should contains both positive and negative residuals (uniform residual plot)	
Regression Equation	$\hat{y} = b_0 + b_1 x$ $b_0 = \bar{y} - b_1\bar{x}$, regression line always pass through $\bar{y}, \bar{x}$ $b_1 = r\frac{s_y}{s_x}$	
Lurking Variable	An unobserved variable that influences the association between variable of primary interest	Confounding effect
Common response	Similar to confounding, difference in x and y has no relationships	

Probability rules	
False Negative: P(NEG|P)	False Positive: P(POS|A)
Sensitivity: P (POS|P)	Specificity: P (NEG|A)
P(Outcome occurring as first n events) = P(outcome as last n events)	
Conditional Probability	
Addition rule for disjoint events $P(A \text{ or } B) = P(A) + P(B)$ If events are disjoint, then events are dependent (if dice is odd, the chance of dice is even can never happen)	General Addition rule $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Multiplication Rule for independent events $P(A \text{ and } B) = P(A)\,P(B)$	Complement $P(A^c) = 1 - P(A)$
Conditional probabilities $P(A\|B) = \frac{P(A \text{ and } B)}{P(A)}$	Intersection $P(A \text{ and } B \text{ and } C) = P(A)P(B\|A)P(C\|A \text{ and } B)$
Independence $P(B\|A) = P(B)$	

Calculator function	**Implication**
Normal Calculation	$P(-x \leq X \leq x)$
Binomial Calculation	$P(X = x), P(X \leq x)$

Standard Normal	$\frac{x-\mu}{\sigma} \sim N\,(0,\,1)$	1. Unimodal 2. Bell shaped 3. Symmetric
Normal Proximation	Binomial distribution Sample Proportion	$np \geq 10$ $n(1-p) \geq 10$
Continuity Correction (cc)	Required for Binomial approximation to normal distribution	$P(X \geq x) \rightarrow -0.5, P(X > x) \rightarrow +0.5$ $P(X \leq x) \rightarrow -0.5, P(X < x) \rightarrow +0.5$

Ling – ST1131 cheatsheet

Distribution	**Usage**	**Mean**	**Standard deviation**	**Hypothesis Testing**	**Standard Error/ Sample Size**	**Confidence Interval** *Point estimate ± margin of error*(m)	**Remarks**
			Population	*Only if given population parameters*			
Normal $X \sim N(\mu, \sigma)$	Known population s.d.	μ	σ	$Z = \frac{X - \mu_0}{\sigma}$	*Known true parameters*	*Known true parameters*	Must be Normal
Binomial $X \sim B(n, p)$	Known pop. probability Two outcomes	$\mu = np$	$\sigma = \sqrt{np(1-p)}$	$Z = \frac{\mu - \mu_0}{\sqrt{np(1-p)}}$ [extreme cases] $P(X \geq x)\ or\ P(x \leq X)$	*Known true parameters*	*Known true parameters*	Approx. Normal w. condition
			Sample	*Must be Normal, or large sample through CLT*			
Sample proportion $\hat{p}$	Unknown pop. probability	$\hat{p} = \frac{X}{n}$	$\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$n = \frac{{z^*}^2}{m}\ p^*(1-p^*)$	$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Approx. Normal w. condition
Sample Mean $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$	Normal distribution or CLT	μ	$\frac{\sigma}{\sqrt{n}}$	$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$n = \frac{z^* \sigma^2}{m}$	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	
Student's T	1. Unknown population s.d. 2. Small sample size	μ	$\frac{s}{\sqrt{n}}$	$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$n = \frac{t^* s^2}{m}$	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	df = (n-1) Degree of freedom: $\lim_{(n-1)\to\infty} t = N$
			Two samples	*Must be Normal, or large sample through CLT*			
Matched Paired	Before & After $\mu_{before} - \mu_{after}$	μ_d	$\frac{s_d}{\sqrt{n}}$	$t_{n-1} = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$	$n = \frac{t^* {s_d}^2}{m}$	$\bar{x} \pm t^* \frac{s_d}{\sqrt{n}}$	df = (n-1)
2 Samples Mean $\mu_1 - \mu_2$	Z test, Compare Differences	$\mu_1 - \mu_2$	$s = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$	$\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$	
2 Samples Mean $\mu_1 - \mu_2$	t test, Compare Differences	$\mu_1 - \mu_2$	$s = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$	$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$	$\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$	Must be between two independent samples df: $\min(n_1, n_2) - 1$
Pooled Test $\mu_1 - \mu_2$	$\frac{{\sigma_x}^2}{{\sigma_y}^2} < 3$	$\mu_1 - \mu_2$	$s = \sqrt{\frac{(n_1-1){s_1}^2 + (n_2-1){s_2}^2}{n_1+n_2-2}}$	$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	df = $n_1 + n_2 - 2$
Two sample proportions $\hat{p}_1 - \hat{p}_2$	$\hat{p}$, Compare Differences	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot SE$	$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$

Ling – ST1131 cheatsheet