Statistics

ML / Stats - Regression Analysis

- **Coefficients** β_i
    - Holding all other variables constant, 1 unit change of x_i averagely leads to β_i units of change in $\hat{y}$
- **P value**
    - Null Hypothesis: $\beta_i = 0$ → 假设这个variable 对 Y 没有影响
    - If P-value ≤ 0.05 (0.01) → Reject Null and favor that $\beta_i \neq 0$
- **Feature Importance**
    - Coefficient β_i 值的大小 代表该variable对 Y 的影响程度大小
        - **Note: 讨论importance时, variables都需要standardize**
    - Statistical Significance: 确认该variable 的 p-value < 0.05

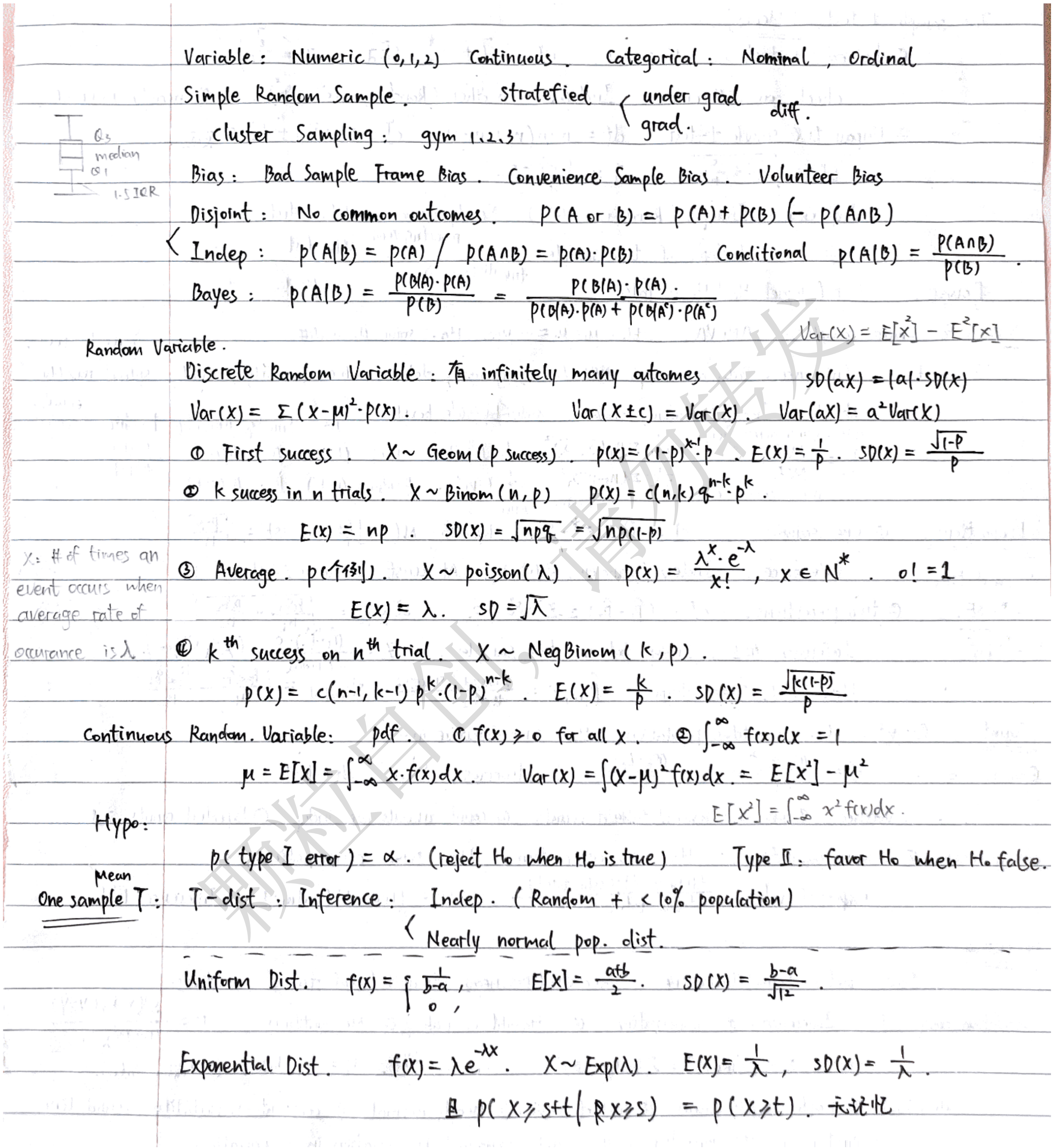
Stats - A/B Testing

- **P Value**: How extreme the observed value is under the null hypothesis
- Law of Large Numbers (LLM)
    - Y_n converges in probability to $\mu_Y \equiv E[Y]$ when sample size n is large
    - https://builtin.com/data-science/law-of-large-numbers
- **Central Limit Theorem** (CLT)
    - https://www.scribbr.com/statistics/central-limit-theorem/
    - CLT allows us to study $\bar{Y}_n$ even if sample data are not normally distributed

Stats - How to choose a test

- 如果是 **Numerical** & **sample mean** 有足够意义
    - 如果 population μ, σ 已知 → Z-test
    - 如果 population μ, σ 未知 → **T-test**
        - 如果treatment 对于个体差异非常大 → paired t-test
            - **https://math.stackexchange.com/questions/1732771/paired-t-test-vs-welchs-t-test**
        - 如果个体差异小
            - https://www.statology.org/paired-vs-unpaired-t-test/
            - 组间差距大 (variance 不同, heteroskedasticity) → **Welch's t-test**
            - 组间差距小 (variance 相同, homoscedasticity) → unpaired (Student's) t-test
        - 如果有超过2组 → ANOVA https://www.qualtrics.com/experience-management/research/anova/
- 如果是 **Categorical** → χ^2 test https://www.investopedia.com/terms/c/chi-square-statistic.asp
- 如果 sample mean 不是讨论对象
    - 用KS Test 来 compare distribution https://towardsdatascience.com/kolmogorov-smirnov-test-84c92fb4158d

- 用 Mann Whitney U Test 来检测median https://datatab.net/tutorial/mann-whitney-u-test



Q3
median
Q1
1.5 IQR

Variable : Numeric (0, 1, 2) Continuous . Categorical : Nominal , Ordinal

Simple Random Sample . Stratefied { under grad / grad . diff .

cluster Sampling : gym 1, 2, 3

Bias : Bad Sample Frame Bias . Convenience Sample Bias . Volunteer Bias

Disjoint : No common outcomes . $p(A \text{ or } B) = p(A) + p(B)$ $(- p(A \cap B))$

{ Indep : $p(A|B) = p(A)$ / $p(A \cap B) = p(A) \cdot p(B)$ Conditional $p(A|B) = \frac{P(A \cap B)}{P(B)}$.

Bayes : $p(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}$

$Var(X) = E[X^2] - E^2[X]$

Random Variable .

Discrete Random Variable : 有 infinitely many atcomes $SD(aX) = |a| \cdot SD(X)$

$Var(X) = \Sigma (X - \mu)^2 \cdot p(X)$. $Var(X \pm c) = Var(X)$. $Var(aX) = a^2 Var(X)$

① First success . $X \sim Geom(p \text{ success})$. $p(x) = (1-p)^{x-1} \cdot p$. $E(X) = \frac{1}{p}$. $SD(X) = \frac{\sqrt{1-p}}{p}$

② k success in n trials . $X \sim Binom(n, p)$ $p(X) = c(n,k) q^{n-k} \cdot p^k$.

$E(X) = np$. $SD(X) = \sqrt{npq} = \sqrt{np(1-p)}$

X: # of times an event occurs when average rate of occurance is λ

③ Average . p(个例) . $X \sim poisson(\lambda)$. $p(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$, $x \in N^*$. $0! = 1$

$E(X) = \lambda$. $SD = \sqrt{\lambda}$

④ k^{th} success on n^{th} trial . $X \sim NegBinom(k, p)$.

$p(X) = c(n-1, k-1) p^k \cdot (1-p)^{n-k}$. $E(X) = \frac{k}{p}$. $SD(X) = \frac{\sqrt{k(1-p)}}{p}$

Continuous Random. Variable: pdf . ① $f(x) \geq 0$ for all x . ② $\int_{-\infty}^{\infty} f(x) dx = 1$

$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$. $Var(X) = \int (x - \mu)^2 f(x) dx = E[X^2] - \mu^2$

$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx$.

Hypo:

$p(\text{type I error}) = \alpha$. (reject H_0 when H_0 is true) Type II : favor H_0 when H_0 false.

Mean
One sample T : T-dist : Inference : { Indep . (Random + < 10% population) / Nearly normal pop. dist.

Uniform Dist. $f(x) = \begin{cases} \frac{1}{b-a} , \\ 0 , \end{cases}$ $E[X] = \frac{a+b}{2}$. $SD(X) = \frac{b-a}{\sqrt{12}}$.

Exponential Dist . $f(x) = \lambda e^{-\lambda x}$. $X \sim Exp(\lambda)$. $E(X) = \frac{1}{\lambda}$, $SD(X) = \frac{1}{\lambda}$.

且 $p(X \geq s+t \mid 且 X \geq s) = p(X \geq t)$. 无记忆

Regression Inference. $SE(b_1) = \frac{S_e \text{ (residual)}}{S_x \cdot \sqrt{n-1}}$. hist: t_{n-2}.

cI: estimate $\pm t^*_{n-2} \cdot SE(b_1)$ (estimate $= b_1$)

	Estimate	std Error	t-value
Intercept			
▭	k	A	k/A.

$H_0: \beta = 0$. $H_A: \beta \neq 0 \longrightarrow$ an association.

Two sample t-test. (Means).

① Paired t-test. $df = n-1$. cI: $\bar{d} \pm t^*_{df} \cdot SE_{\bar{d}}$. $T = \frac{\bar{d}-0}{SE_{\bar{d}}}$.

check on differences: Indep of each other (Random, <10% pop). Diff. nearly normal.

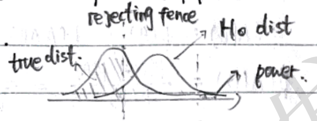
② Unpair two sample t-test $df = \min(n_1-1, n_2-1)$. cI: $(\bar{x}_1-\bar{x}_2) \pm t^*_{df} SE_{\bar{x}_1-\bar{x}_2}$.

$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ $T = \frac{(\bar{x}_1-\bar{x}_2)-0}{SE}$

Inference: Indep. (Random, < 10%). Nearly normal population dist.

Indep. of two samples.

Power: p(reject H_0 | H_0 is false).



Means for 3 more samples: ANOVA. $H_0: \mu_1 = \mu_2 = \cdots \mu_k$. H_A: some μ_i is diff.

F-dist: Inference: Data are indep. within/across groups; data in each group nearly normal; spread in each group roughly equal.

check: Indep (random, <10%). side-by-side boxplot

mean square across groups

$F = \frac{MSG}{MSE} = \frac{\frac{1}{k-1}\Sigma n_i(\bar{x}_i-\bar{x})^2}{\frac{1}{n-k}\Sigma(n_i-1)^2 s_i^2}$ K组数据 每组n个.

mean square error

	Df	Sum Sq	Mean Sq	F-value	P value
year	(k-1)	A	A/(k-1)	$\frac{A/(k-1)}{B/(n-k)}$	...
Residuals	(n-k)	B	B/(n-k)		

Proportions. ① one sample. cI: $\hat{p} \pm z^* \cdot SE$. hypo 看 $N(p, \sqrt{\frac{pq}{n}})$. $SE = \sqrt{\frac{p \cdot q}{n}}$

margin of error

Inference: Indep (random, < 10%). At Least 10 success / failures

Margin of Error: $z^* \cdot SE$. $= z^* \cdot \sqrt{\frac{pq}{n}}$

② two proportions. cI: $(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot SE_{\hat{p}_1-\hat{p}_2}$. $SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

Inference 同上, Indep. between two samples. $z = \frac{(\hat{p}_1-\hat{p}_2)-0}{SE_{pooled}}$ (只有 SE_{pool}).

$H_0: p_1 = p_2 \implies$ hypo: $\hat{p}_{pooled} = \frac{\text{Success 1} + \text{Success 2}}{n_1+n_2}$. $SE_{pooled} = \sqrt{\frac{\hat{p}_{pool}\hat{q}}{n_1} + \frac{\hat{p}_{pool}\hat{q}_{pool}}{n_2}}$

Counts. ① χ^2: H_0: ⋯ matches population, H_A: some difference in⋯.

Goodness-of-Fit. $\chi^2 = \frac{(O_1-E_1)^2}{E_1} + \cdots + \frac{(O_k-E_k)^2}{E_k}$, k categories, χ^2_{k-1}. $df = k-1$

生日、基因型

Inference: ① one-dimensional table of counts ② counts in cells are indep ③ Expected counts ≥ 5

Test for indep: ② 2 more populations. H_0: no association, H_A: some an association.

expected value: $\frac{\text{(row total)} \cdot \text{(colm total)}}{\text{table total}}$ $df = (\text{row}-1)(\text{colm}-1)$ Inference 同上

Regression.

Scatter plot: Form (lin. curve). Direction (pos. neg) Strength (weak. strong) Outliers

correlation coeff: R: Inference: ① quantitative ② straight enough ③ No outliers ④ Indep data. $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1) \cdot S_x \cdot S_y}$

error: $\Sigma(\text{residuals})^2 = \Sigma(\text{obs}-\text{exp})^2$ $b_1 = r \cdot \frac{S_y}{S_x}$. $b_0 = \bar{y} - b_1\bar{x}$.

Inference: ① graph roughly linear ② hist of residuals nearly normal ③ constant Variability around line

r^2 proportion of the variation in y-variable explained by variation in x-variable.

Stats - 异动归因

- 问题类型: 某metric突然变化，如何分析？
- 首先确认问题**真实性**
    - 读取data，存储新data 是否出错？(e.g. 新上线商品分类错误、计量错误)
    - 问题原本是否有，突然的提升 是否statistically significant ?
- 罗列因素
    - **外部因素**: PEST
        - Policy, Economics, Social, Technology
    - **内部因素**: index拆解
        - 公式分解
        - 维度考虑，e.g. gender, age, operation system, vertical
        - 比较前后的**distribution是否相同**
            - $E[Y] = \sum E[Y_i|X_i] \cdot P(X_i)$, Y_i 变化 & $P(X_i)$ 分布变化 都有影响
            - **Simpson's Paradox**: https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765
            - 解决方案
                - **Segregate the data in groups** or **aggregate data together**
                - https://towardsdatascience.com/simpsons-paradox-how-to-prove-two-opposite-arguments-using-one-dataset-1c9c917f5ff9