STATISTICS

AP STATISTICS

美国大学先修课统计学

作者：梁梓涵

Author: Zihan Liang

Unit 0 Introduction

1. Individual, Variable 和 distribution 的意义

个体 Individual：n & N（一组数据描述的对象）

变量 Variable：性质（个体的任何特征）

分布 Distribution：Variable 值+(How often) Frequency 变量取哪些值+取各个值的频率

Unit 1 Exploring Data: Describing Patterns and Departures from Patterns (20%-30%)

A. Exploring Categorical Data

1. 表格 one-way and two way table; Frequency and relative frequency 的区别和使用

a. <u>One-way table 和 Two way table 的区别</u>

Variable	Frequency=Counts
A	1
B	2
C	3
total	6

One-way table

Var 1 \ Var2	Female	Male	Total
Almost no	96	98	194
Some chance	426	286	712
50-50 chance	696	726	1416
Good chance	663	758	1421
Almost yes	486	597	1083
Total	2367	2459	4826

Two-way table

One-way table 可以使用 frequency (frequency table)或者 relative frequency (relative frequency table)

Two-way table 一般不用 relative frequency

One-way table 只有一个变量；Two-way table 有两个 categorical variable

b. <u>Frequency and relative frequency 的区别和使用</u>

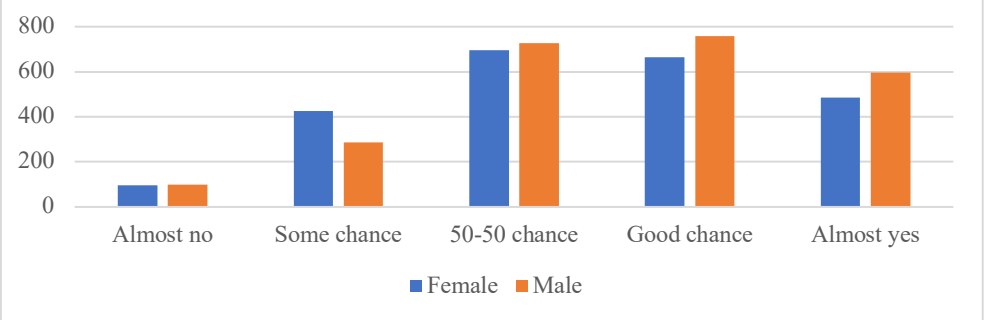
Frequency=counts（次数）；Relative frequency=Freq./total (percent)

*注意：百分比相同，n 不一定相同

2. Marginal distribution for two-way table 的计算和画图比较 bar (side-by-side bar graph)

$$\text{Marginal distribution} = \frac{\text{行 total 或列 total}}{\text{总 total}}$$

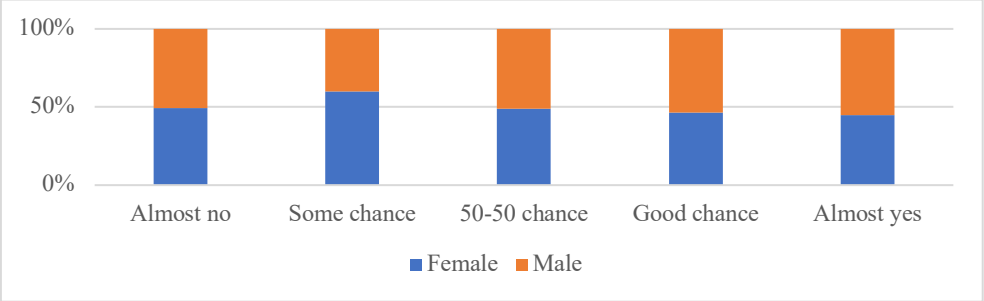
以上述 two-way table 为例，almost no chance=194/4826=4.0%, some chance=712/4826=14.8%



3. Conditional relative frequency 的计算和画图比较 segmented-bar (segmented bar graph)

$$\text{Conditional distribution} = \frac{\text{交叉值}}{\text{condition 的行/列 total}}$$

以上述 two-way table 为例，女生中 almost yes 的人数=486/2367=20.5%



4. 判断 association 根据 conditional distribution 判断 categorical variable 相关性（独立性）

Association---相关性=related

No association---不相关=independent

相关性 Association: Knowing the value of Var A help to know the value of Var B---A and B associated.

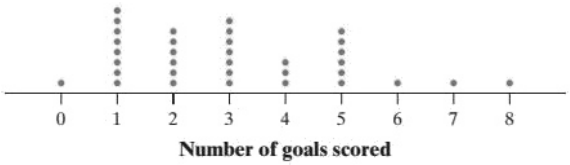
5. Pie chart and bar graph 各自特点

Pie chart 必须包括构成一个总体的所有类别（不是 100%需要加 other，面积加和必须是 100%）

Bar graph 则不需要单独加 other 使其构成 100%

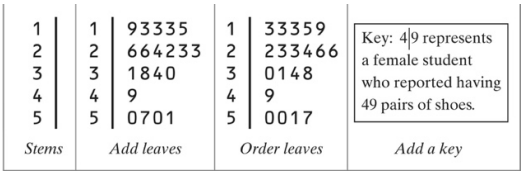
B. Graphical Displays of Quantitative Distributions

1. Dot plot 的特点



X 轴成比例；Exact value（能查出值）；Small data set（数据少）；能看出 shape

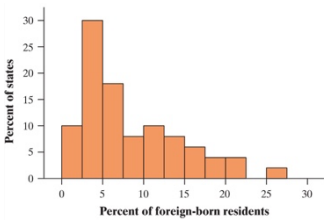
2. Stem plot 的特点



Key! ；Handful 原则（stem5-10 个）；Exact values（能查出值）；Small data set（数据少）；能看出 shape

3. Histogram 的特点及做法

a. <u>Histogram 的特点</u>



无具体数值；成比例；large number of observations（数据多）；能看出 shape

b. Histogram 做法

Step 1: Divide the data into classes of equal width.

0–5；5–10；10–15；15–20；20–25；25–30

注意边界值，例如 5，10，15 的去处

0 to <5；5 to <10；10 to <15；15 to <20；20 to <25；25 to <30

Step 2: Find the count (frequency) or percent (relative frequency) of individuals in each class.

Step 3: Label and scale your axes and draw the histogram.

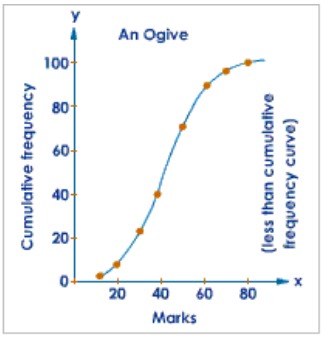
c. Histogram 和 Bar chart 区别

区别一：Histogram 的各矩形连续排列（如果出现空白则代表相应 class 没有数据，因此空白具有数学意义）；Bar chart 的各矩形不连续分布，中间的空格没有数学意义

区别二：Histogram 用来描述 quantitative variable 中的连续变量；Bar chart 用来描述 Categorical variable 或者离散型的 quantitative variable

区别三：Histogram 各矩形的宽度代表每组的组距，Bar chart 的宽度没有数学意义

4. Cumulative frequency plot 的阅读



y 轴：cumulative frequency=percentile

x 轴：变量

C. 会在图中识别/计算以及线性转化对以下值的影响

1. Measuring center: median, mean

a. 平均数 Mean

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

b. 中位数 Median

中位数 M 是分布的中点

1) 从小到大排列所有观测值 Observations。

2）如果观察数 n 是奇数 odd，中位数 M 是最中间的数字。

3）如果观察数 n 是偶数 even，中位数 M 是有序列表中两个数字的平均值。

c. Mean, median and outlier

Mean **not resistant** to extreme values/outliers/skew

Medium **resistant** to extreme values/outliers/skew strongly skewed 或 outliers 多选择 median

2. Measuring spread: range, interquartile-range, standard deviation

a. 范围、最大值和最小值 Range, Maximum and Minimum

Maximum：所有观察值中最大的数字；Minimum：所有观察值中最小的数字

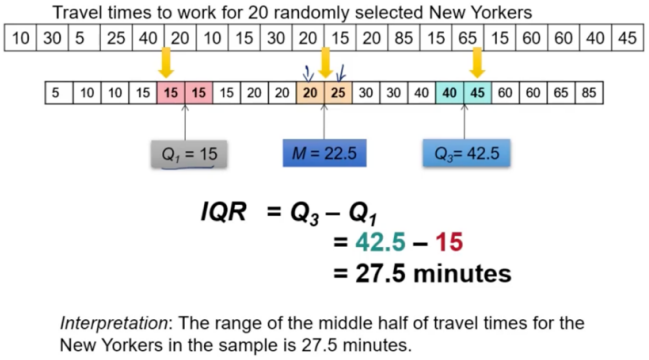
$$Range = Maximum - Minimum$$

b. 四分位点与四分位距 Quartiles & Interquartile Range (IQR)

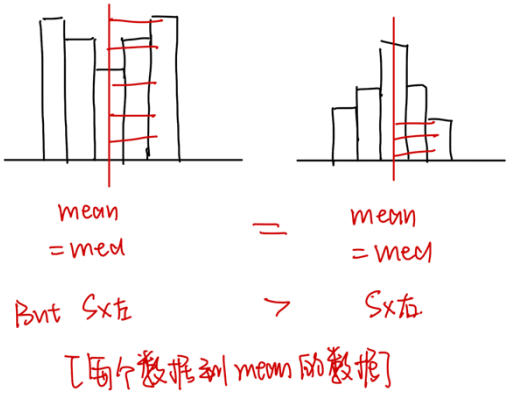
1) 按递增顺序排列观察值并找到中位数 median M。
2) 第一个四分位数 first quartile Q1，是位于有序列表中中位数左侧的观察值的中位数。
3) 第三个四分位数 third quartile Q3，是位于有序列表中中位数右侧的观测值的中位数。

四分位距 Interquartile Range (IQR) 定义为：

$$IQR = Q_3 - Q_1$$



c. Standard deviation



Standard deviation is the **typical distance** from data to mean.

Sample 的 SD：

$$S = \sqrt{\frac{\sum(x_i - x)^2}{n-1}}$$

d. Center, spread and outlier

Mean 和 standard deviation 搭配

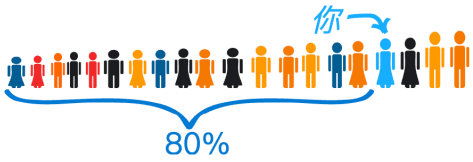
Medium 和 IQR 搭配

3. Measuring position: quartile, percentiles, standardized scores (z-scores)

a. Percentile

The value below which a percentage of data falls. 低于该值的数据的百分比

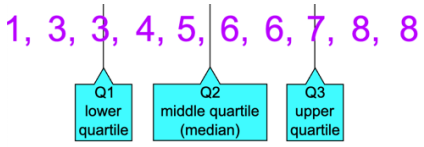
Example：你是班上 20 个学生里身高第四的学生，80%的学生比你矮：



That means you are at the 80th percentile.
如果你的身高是 1.85m，"1.85m" 是班上身高的 80% 百分位数。

b. Quartile

另外一个相似的概念是四分位数 quartile，它把数据分成四份：
Example: 1、3、3、4、5、6、6、7、8、8



Quartile 1 (Q1) = 3 ---25th percentile
Quartile 2 (Q2) = 5.5 --- 50th percentile (median) 50%面积 **equal-area point**
Quartile 3 (Q3) = 7 --- 75th percentile
Mean --- **balance point**

c. Standardized scores (z-scores)

$$standardized\ score: z = \frac{x - mean}{standard\ deviation}$$

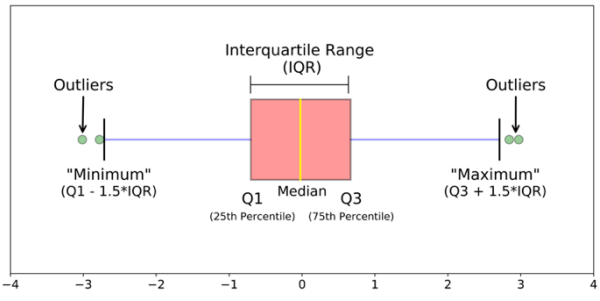
Interpretation: 该 x 距离 mean 有 z 个 SD

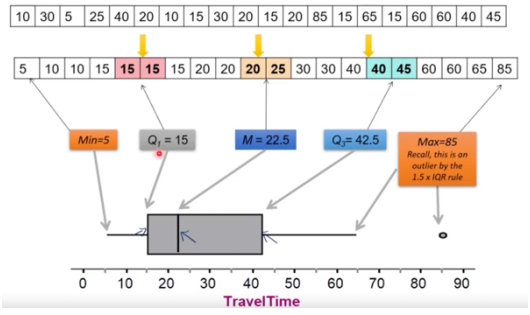
$$z > 0 \rightarrow x > \text{mean z 个 SD}$$
$$z < 0 \rightarrow x < \text{mean z 个 SD}$$
$$z = 0 \rightarrow x = \text{mean}$$
$$z > 2 \rightarrow \text{outliers}$$

4. Boxplots 的特点

a. 五参数总结法 Five number summary

五参数总结法包括 the smallest observation、the first quartile、median、the third quartile 和 the largest observation，按从小到大的顺序排列。





Boxplots 无法看出 shape

b. 离群值 Outliers

在数据中有一个或几个数值与其他数值相比差异较大。

方法一：$Q_1 - 1.5 \times IQR$以及$Q_3 + 1.5 \times IQR$构成一个离群值的范围：凡是在此范围之外的均为离群值 Outliers。
方法二：$mean \pm 2 \times Standard\ deviation$构成一个离群值的范围。

5. The effect of changing units on summary measures (Transforming data)

a. <u>Adding or subtracting a constant ±a (shape不变)</u>

mean, median, quartiles and percentiles ± a
observations ± a
spread 不变 (ranges, IQR, SD)

b. <u>Multiplying or dividing a constant × & ÷ b</u>

mean, median, quartiles and percentiles × & ÷ b
observations × & ÷ b
spread (ranges, IQR, SD) × & ÷ b

D. Comparing Distributions of Quantitative Variable

1. Comparing center and spread

Center: The median XXX of A (MED_A=___unit) is greater than B (MED_B=___unit).
Spread: The range of XXX of A (Range A=___unit) is wider than B (Range B=___unit).

2. Comparing clusters and gaps

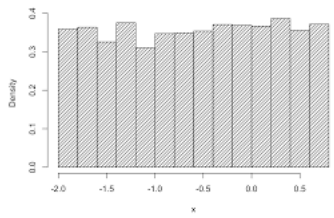
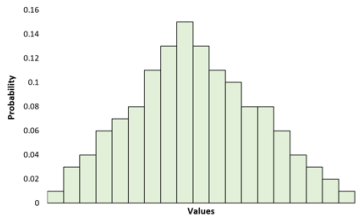
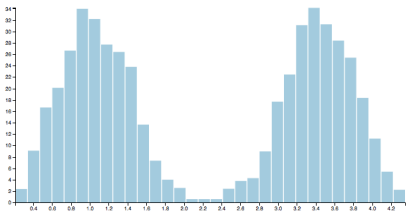
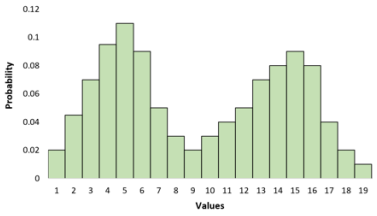
Cluster: A group of values sticks together away from other groups.
Gaps: The ‘’large’’ open space between some data points.

3. Comparing outliers and other unusual features

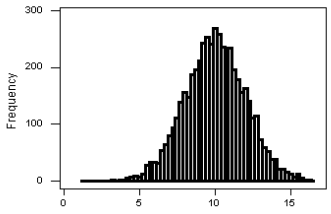
Outliers: There are ___ (potential) outliers for A/B, whereas no obvious outliers for A/B.

4. Comparing the shapes

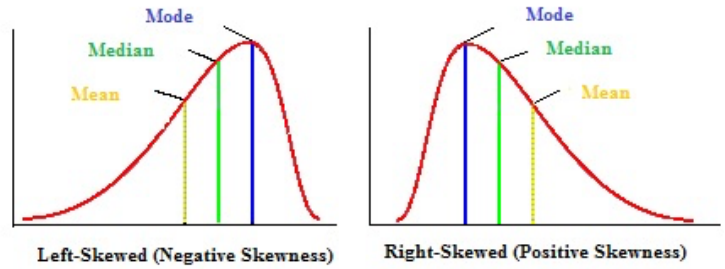
Shape: Both XXX of A and B are _____(right/left-skewed & whereas…)

	
Uniform	Unimodel/mound-shaped
	
Two clusters with a gap	Bimodel

对称性 Symmetry



Roughly symmetric



Left-skewed（左脚）；Right-skewed（右脚）

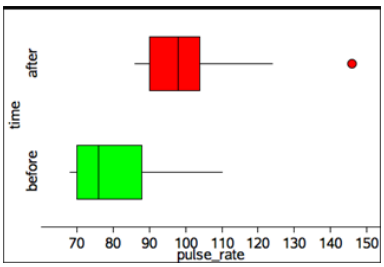
在偏态分布中，mean 会往尾巴方向跑

5. Parallel boxplot, back-to-back stemplot and side-by-side histogram

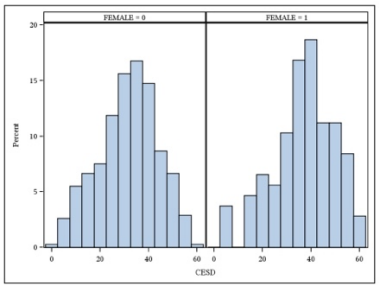
Females		Males
	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

Key: 2|2 represents a male student with 22 pairs of shoes.

back-to-back stemplot



parallel boxplot

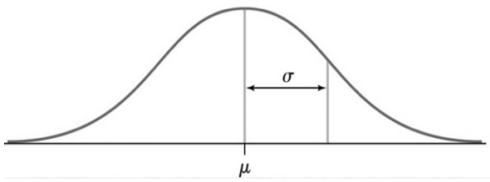


side-by-side histogram

E. Normal Distribution

1. Properties of the normal distribution

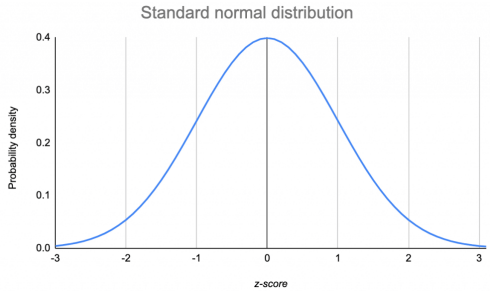
a. Normal distribution



Shape: symmetric, single-peaked, bell-shaped；Center: mean=median；Spread: σ

正态分布的两个参数为 mean 和 SD，简写为 $N^{(\mu,\sigma)}$，其中μ决定函数的位置，σ决定其形状（σ越大，曲面越平缓，σ越小，曲面越陡峭）。

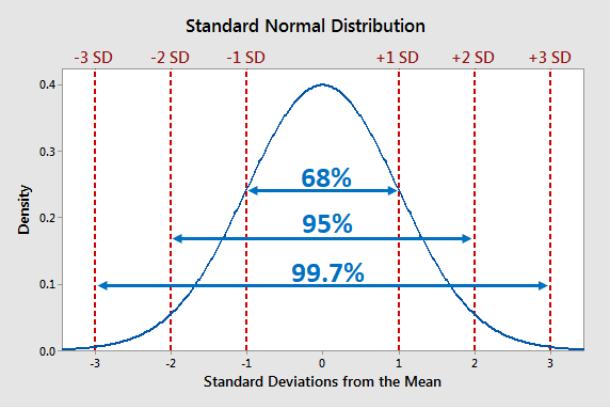
b. Standard normal distribution



当 center 变为 0，spread 变为 1 时，为标准正态分布。任何一个正态分布都可以通过 z-score 转化为

标准正态分布。

2. Empirical rule



68% of the observations fall within σ of μ.

95% of the observations fall within 2σ of μ. ($2 \times SD$ rule of the outliers)

99.7% of the observations fall within 3σ of μ.

3. Normal distribution calculation

a. <u>How to find areas in any normal distribution</u>

Step 1: State the distribution and values of interest. 画图. N(μ,σ)，根据题意，画出 area 和 x(z 值)

Step 2: Calculation (Two methods to perform calculation).

- Compute a z-score for each boundary value and use Table A or technology to find the desired area under the standard Normal curve.
- Use the normalcdf command and label each of the inputs.
  - 计算器 MENU 2---DIST---NORM---Ncd
  - Data 选择 Variable
  - 输入相关数值 Normalcdf (lower=____ upper=____ μ=____ σ=____)
  - 注意：正负无穷输 100 或 1000；默认使用 Z 值，$\mu = 0$，$\sigma = 1$

Step 3: Answer the question.

b. <u>从面积求 Z 值</u>

Step 1: 计算器 MENU2---DIST---NORM---InvN

Step 2: Data 选择 Variable

Step 3: 输入相关数值 invNorm (central/left/right area=___ μ=____ σ=____)

注意：如何判断 area 类型

Unit 2 Exploring Bivariate Data (5-7%)

1. Analyzing patterns in scatterplots

a. DFSO

There is a (extremely/moderately) (strong/weak), (positive/negative)-association, (linear/nonlinear) relationship between __x__ and __y__. There is no obvious outliers/There is/are _____ potential outliers.

b. The interpretation of the strength

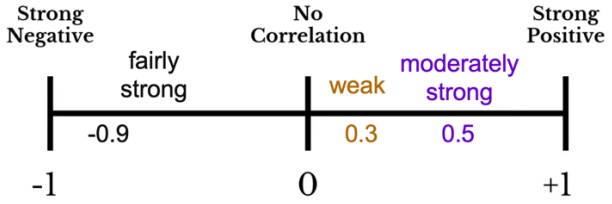
Strength 代表的是点与点之间的线性关系是否清晰，可以用 correlation (r) coefficient 来衡量 strength。

2. Correlation coefficient (r)

a. Correlation coefficient 的定义与计算

Correlation (r) coefficient：测量 linear relationship 的 strength 和 direction

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_y}\right) = \frac{1}{n-1}\sum(Z_x)(Z_y)$$



r 介于[-1,1]之间，$r > 0$代表 position-association，$r < 0$代表 negative-association

b. Correlation coefficient 的九大注意事项

x 与 y 互换，r 不受影响，但斜率会被影响
x 与 y 改变 unit，r 不受影响
r 值无单位，但斜率有单位
r=-1 不可以保证一定是 negative linear，要看 residual plot，因为有可能不是 linear
r 值不能代表 non-linear 强弱，其不代表曲线(curve)相关
correlation 不等于 causation，有相关性不一定是因果关系
categorical variable 不能计算 r 值
r 值不足以描绘一个双变量关系

c. 通过 R-sq 算 r 值

$$r = \sqrt{r^2\ (R-sq)}$$（正负要看 slope）

3. 构建 Least-square regression line（解读统计软件的结果）

a. Interpreting the regression line

$$\hat{y} = a + bx$$

$\hat{y}$: predicted value of y for a given x
b: slope
a: y-intercept

b. 解读统计软件的结果

	Minitab	
Predictor	coef	
Constant	38267	← y-intercept (a)
Miles driven (x)	-0.16292	← slope

$$\widehat{price} = 38267 - 0.16292\ (miles\ driven)$$

注意：x 和 y 的解释（建议直接标注在公式上）；predicted 标

4. 解释 slope, y-intercept, S, R-sq 的含义

a. Interpreting the slope

When (the name of x) increase 1 (unit), predicted (the name of y) will increase/decrease (sbsolute value of slope) (unit).

Example: When miles driven increase 1 thousand, predicted price will decrease 0.16292 dollar.

b. Intrerpreting the y-intercept

When (the name of x) is equal to 0 (unit), predicted (the name of y) is equal to (y-intercept) (unit).

Example: When miles driven is equal to 0 thousand, predicted price is equal to 38267 dollars.

c. Interpreting the residual and S (standard deviation of the residual)

Residual: When (the name of x) was ____(unit), the predicted (the name of y) was ____(unit) lower/higher than the actual y.

S (Standard deviation of the residual): When using ____ to predict ____, we would expect a typical residual to be (S%).

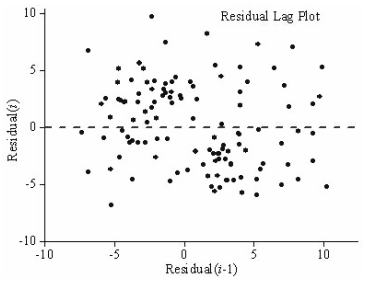
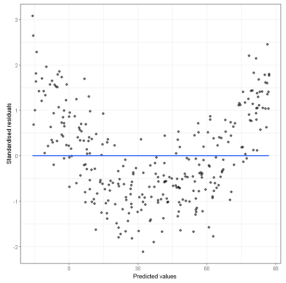
d. Interpreting the R-sq

There are ____% of the total variation in (the name of y) be explained by the LSRL of using (the name of x) to predict (the name of y).

5. 评价模型的好坏&判断是否 Linear

模型好坏判断：通过 S 和 R-sq 来判断，当 S 越小（residual 越小代表误差越小），R-sp 越大（被 explained 的 variation 越多）时，线性模型越好。

Linear 与否判断：在 Residual plot 中，散点越 random，越 linear。

	
Random residual plot---linear	Curve（或其他） residual plot--- nonlinear

6. Two types of influential points

High-leverage points：x 极大或极小值，远离 x 的 mean，影响 slope（low-leverage points 不影响）

Y 方向上的 outliers：residual 大，对 correlation 的影响大

Unit 3 Sampling and Experimentation: Planning and Conducting a Study (10-15%)

1. 四种方法的区分

Census 普查	Sample survey 抽样调查 （是一种 obs. study，单变量）
收集 population 中每个 individual 的数据	收集 sample (n)去预测 population (N)
得到 population 数据 (μ, σ)	得到 sample 数据 $(\bar{x}, s)$

Observational study	Experiment
观察测量 variable，不试图影响（只观察）	给每个个体实施 treatment，去测量 response
Bivariate 双变量	Bivariate 双变量
No cause and effect	研究 cause and effect
/	有 random assign

2. 区分和设计 random sampling

a. Simple Random Sample (SRS)

Every group of n individuals in the population has an equal chance to be selected as the sample. (n↑ sampling variability↓)

Sampling method:

Step 1: label---给 population (N) 中的每个 individual 一个特定的数字标记（从 1 到 N）。

Step 2: randomize---用 random number generator，在 1 到 N 之间，产生 n 个整数 (Sample 1)。

局限性：要给每个 population 一个特定的数字标记；当 N 过大，SRS 使用困难

b. Stratified Random Sample

The population is divided into groups called strata and a simple random sample is selected from each stratum. It is homogeneous group. 需要把 population 根据特征分成组，每组内随机抽样。（组不同 组内相同）

Sampling method:

Step 1: using xxx as strata---根据某些特征进行分层，在每层按比例进行抽样

Step 2: label all individuals within each stratum. Randomly draw X different integers from each stratum.

Example:

取 5 人	← 20 岁 (strata 1)
取 10 人	← 30 岁 (strata 2)
取 15 人	← 40 岁 (strata 3)

Using all resulting 30 people accordingly as the Stratified Random Sample.

c. Clustered Random Sample

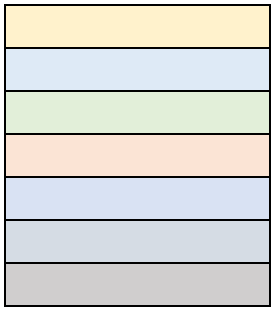
In cluster sampling, the population is divided into heterogeneous groups called clusters, and we then take a random sample from among all the clusters. 把 population 任意分成完全一样的群。（组相同 组内不同）

Sampling method:

Step 1: using xxx as cluster.---随机分成个数完全一样的 cluster

Step 2: random---用 random number generator 从所有 clusters 中任意选择一个，使用该 cluster 中所有 individual 作为样本。

Example:



The cluster 7 is selected as samples using random number generator.

d. Systematic Random Sample

Systematic random sampling is a method to select samples at a particular preset interval.

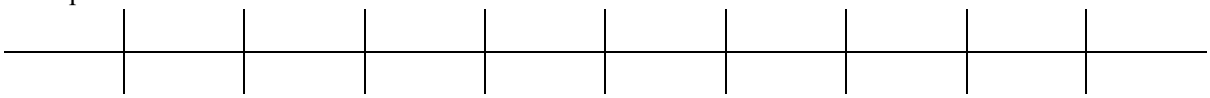
Sampling method:

Step 1: 将总体中的所有个体按照一定顺序排序

Step 2: 确定起点 randomly

Step 3: 根据比例等距取样

Example:



Random：在 1500 个人的 population 中首先 Random 在 1-50 中确定起点（起点之前不要了）

Select：然后每 50 个(等距)取 1 个样，取 30 次（n），得到 30 个 different number

使用最后对应的 30 人作为一个 n=30 的 Systematic Random sample

3. Sources of bias in sampling and surveys

a. Poorly designed sample (non-random sample)

Convenience sample（方便调查者，街头问卷调查） + voluntary response sample（发问卷，志愿填）产生的 bias，因为其**没有 random**。

b. Undercoverage bias

当总体的一部分子集被忽略掉导致的偏差。

比如用座机 landline number 作为 population list 来抽 random sample 时，忽略了没有座机的人群。

c. Non-response bias

由于样本中的个体拒绝参加或无法参加调查导致的偏差。其通常发生在 design random sample 后**调查过程中**。

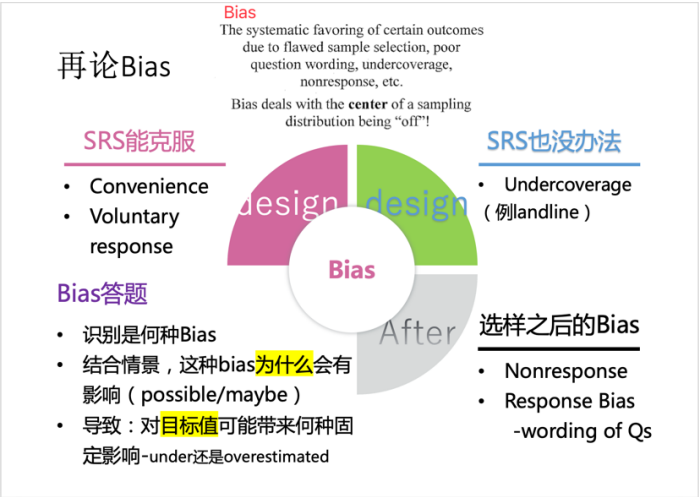
比如抽取了 1500 人的 random sample，白天通过电话调查失业率。接听电话人为 300 人，其中约 180 人没有工作。而实际该市大约只有 20%失业率。因为未回复的人大约在工作，而接电话的人无法代表未回复的人，因此导致失业率被 overestimated。

d. Response bias

因为 ethical, age 等原因撒谎导致了 incorrect answers。其发生在**调查过程中**，因不恰当的方式导致 individual 撒谎，导致 answers 都固定朝某个答案偏移。

Wording of the question：由于问卷的问题有倾向性引导，导致 answers 都固定朝某个答案偏移

e. 再论 bias



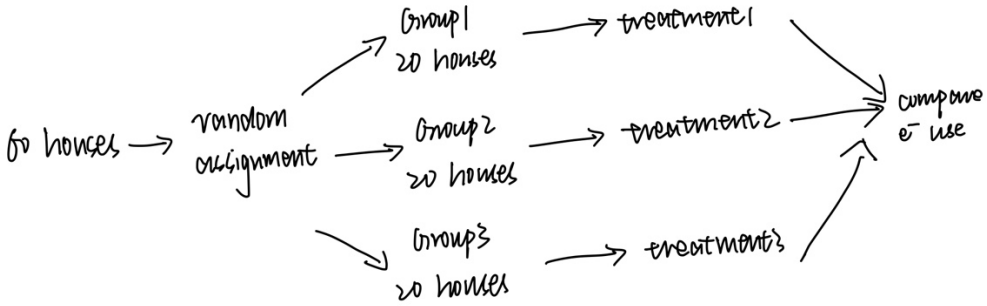
4. 区分+设计 completely randomized design/block randomized design/matched pair design

a. Completely randomized design

用 chance 分配 experiment units 去对应 treatment，确保机会均等、数量均量

Example:

To implement the design, state by labeling each house with a distinct number from 1 to 60. Write the labels on 60 identical slips of paper, put them in a hat, and mix them well. Draw out 20 slips. The corresponding homes will be given digital displays showing current electricity use. Now draw out 20 more slips. Those homes will use a chart. The remaining 20 houses will be given information about energy consumption but no way to monitor their usage. At the end of the year, compare how much electricity was used by the homes in the three groups.



模版总结：To implement the design, state by labeling each XXX with a distinct number from 1 to X. Write the labels on XX identical slips of paper, put them in a hat, and mix them well. 描述如何展示结果。Draw out X slips（抽取第一组）.说明第一组 treatment。Draw out X slips（抽取第二组）.说明第二组 treatment。Draw out X slips（抽取第三组）.说明第三组 treatment。说明实验目的。

b. Block randomized design

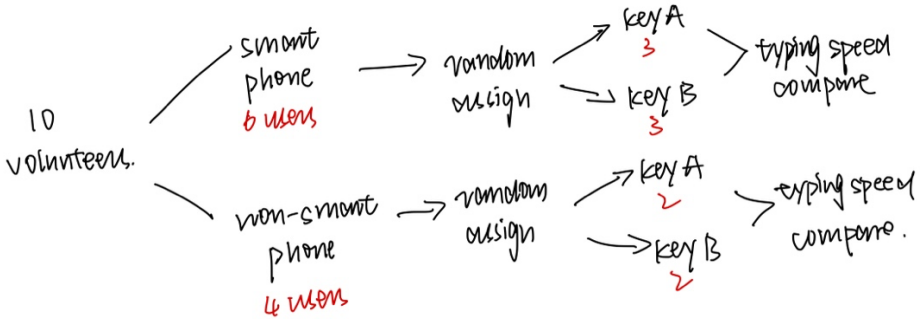
先把 exp. Unit 分成 blocks（根据已知性质分 男/女 新/旧），再各个 block 内使用 completely randomized design。

优点：降低 variation，提供更多信息，一种特殊的 block randomized design 是 matched-pair design。

Example:

To implement the design, state by labeling each volunteer with a distinct number from 1 to 10. Divide the volunteers into two blocks according to whether they have smartphones or not. In "smart phone users" block,

let 6 volunteers stand at random. Randomly select 3 people in “smart phone users” block to use Key A; then let the remaining 3 people use Key B. In “non-smart phone users” let 4 volunteers stand at random. Randomly select 2 people in “smart phone users” block to use Key A; then let the remaining 2 people use Key B. At the end of the assignment, the typing speed of volunteers with a cell phone and using Key A, with a cell phone and using Key B, without a cell phone and using Key A and without a cell phone and using Key B was compared.



c. Matched-pair design

先俩俩配对 exp. Unit（不能 random），在 random assign 配对小组内部，谁先接受哪个 treatment。（本质上是俩俩配对再相减，每个 difference 作为一个 sample）

分析结果时先算 difference 再求 mean。优点：能检测更小的 difference。

With subjects in pairs：这种 pair 中，我们将被试者进行 group，并且每一个人给予不同的 treatment，在一个 pair 中有两个人，其中一个人是 control group（注意不同的 treatment 也可以是 control）。

Example：我们要比较两个清洁剂效果如何，手是我们将 50 个脏盘子分成 25 组，每组脏的程度类似(注意这里不一定是随机)，每一组中随机选择一个用清洁剂 1，另一个用清洁剂 2，然后比较测试结果。

With Subjects as Individuals：这种 pair 中，每一个被试者接受同样的两种 treatments，但次序不同。每一个 individual 自己是自己的 control group。

Example：我们测试一款新的汽车轮胎，我们选择 50 辆两汽车，其中随机选出一半的汽车左轮装新的轮胎，石轮装旧的轮胎；剩下的汽车则恰好相反。然后比较测试结果。

5. Experiment language

Explanatory variable	x:施加的变量	Experiment unit	实验对象
Factors	多个 x（因素）	Subject	人（作为实验对象）
Levels	水平（x 的设定值）	Replication	重复
Response variable	y:测量值		
Treatment	具体自变量组合(trails) Treatment 数量=x数量 × x取值 level		
Confounding variable	混淆变量：与自变量和因变量均相关的变量，confounding variable 使自变量与因变量之间产生了虚假的关系。		
Random assignment	Experiment units are assigned to treatments at random, that is using some sort of chance process. 作用：eliminate 潜在的 confounding variable		

Example: What are the effects of repeated exposure to an advertising message? The answer may depend on both the length of the ad and on how often it is repeated. An experiment investigated this question using 120 undergraduate students who volunteered to participate. All subjects viewed a 40-minute television program

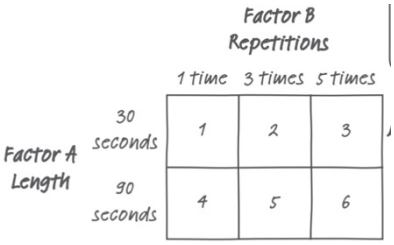
that included ads for a digital camera. Some subjects saw a 30-second commercial; others, a 90-second version. The same commercial was shown either 1, 3, or 5 times during the program. After viewing, all the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it.

Experiment unit (subjects): 120 undergraduate students

Replication: 1, 3, or 5 times

Factors: the length of ads; times of replication

Treatments: (1 time, 30 seconds), (3 times, 30 seconds), (5 times, 30 seconds), (1 time, 90 seconds), (3 times, 90 seconds), (5 times, 90 seconds)



6. Well-designed experiment

- Random assignment: allotment of treatments to experimental units
- Replication $n \geq 2$
- Comparable groups: comparisons of at least 2 groups, one of which could be control group（好但非必要）
- Control potential confounding variable

7. 因果关系 Causation 的判断

重点看 Random assign 而不是是否存在 control group；非 Random assign 只能得出 x 和 y 相关性而没有 Causation

关于结论：没有 bias 的 Random sample 可以上升到 population，而 Volunteer 只能上升到和他们自己 similar 的群体

Unit 4 Probability (10-20%)

1. 给 probability 设计 simulation 思路

Step 1: State 概率问题

在某随机问题中，某结果的出现概率是多少（以此判断是正常还是小概率事件）

Step 2: Plan 编码规则

利用 chance device，注意 record 和 ignore 的数据

Step 3: Do

根据编码规则，重复 30-100 次算出概率，画表格/dotplot/文字

Step 4: Conclude

根据得到的概率，回答问题并作判断

Example：

若一直买彩票中奖的比例是 20:600。如果小明同学如果买了 10 张彩票。请问不中奖的概率是多少？

State：这个例子中，小明不可能不断的买彩票（一次买 10 张，买无数次）。因此我们就可以把这件事模拟成抓阄；

Plan：准备 600 张纸，其中 20 张是蓝色的。每次抽 10 张，记录下蓝色纸的数量；

Do：将上面的行为多次重复。比如重复 1000 次，得到 1000 个数字；

Conclude：记录 1000 个数字中 0 的比例。这就接近于不中奖的概率。

2. Probability 的含义

a. 经典概率

$$\text{经典概率} = \frac{\text{发生过的事件}}{\text{样本空间}}$$

b. 预测概率

Law of large numbers

If we observe more and more repetitions of any chance process, the proportion of times that a specific outcome occurs approaches a single value (the probability of that outcome).

Long-run relative frequency (probability=relative frequency)

The probability of any outcome of a change process is a number between 0 and 1 that describes the proportion of times the outcomes would occur in a very long series of repetitions.

3. Exclusive events and addition rules

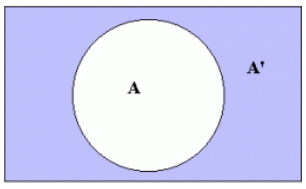
a. Basic probability rules

For any events, $0 \le P(A) \le 1$.

$$P(A) = \frac{number\ of\ outcomes\ corresponding\ to\ Event\ A}{total\ number\ of\ outcomes\ in\ sample\ space\ (universal\ set)}$$

If S is the sample space in a probability model, P(S)=1 (The sum of all probabilities is 1.)

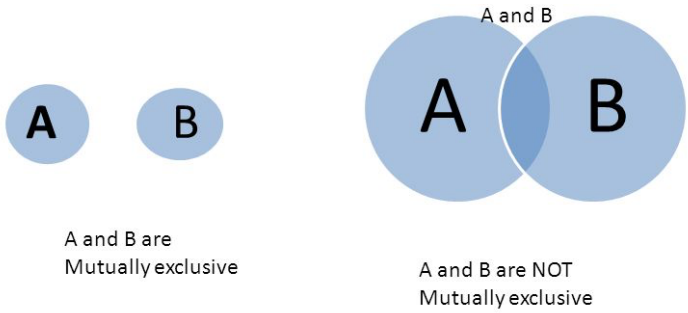
b. Complement set



$P(A') = P(A^c) = 1 - P(A)$

A^c为 A 的补集 Complement set，$P(A^c)$是不发生 A 的概率。

c. Addition rule for mutually exclusive events



$A \cap B$ （A 和 B 同时发生 A and B）

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$（A 发生或 B 发生 A or B）

特殊：当没有交集时 (A and B are mutually exclusive) $P(A \cup B) = P(A) + P(B)$

4. Conditional probability and independence

a. Conditional probability

$$P(B|A) = the\ probability\ that\ event\ B\ happens\ \boldsymbol{given\ that}\ event\ A\ has\ happened.$$

在事件 A 发生的条件下事件 B 的概率 $P(B|A) = \frac{P(A \cap B)}{P(A)}$

在事件 B 发生的条件下事件 A 的概率 $P(A|B) = \frac{P(A \cap B)}{P(B)}$

In the AP Exam, use a verbal equivalent instead of A and B. E.g.,

$$P(reads\ NYT|reads\ USA\ Times) = P(B|A)$$

b. Independent event and general multiplication rule

但一个 event 发生并没有改变另一个 event 将发生的可能性时，两件事独立

If $P(A|B) = P(A)$ and $P(B|A) = P(B)$ → independent

General multiplication rule

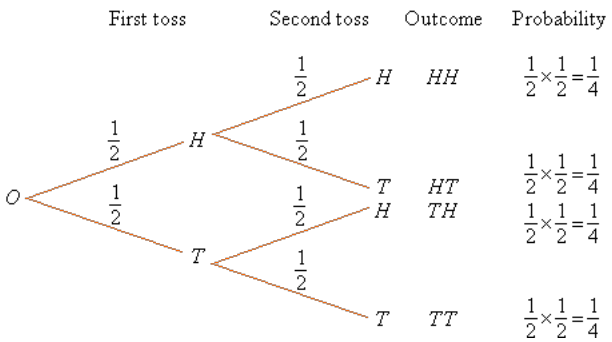
$$P(A \cap B) = P(A) \times P(B|A)$$

where P(B|A) is the conditional probability that event B occurs given that event A has already occurred.

c. Tree diagrams

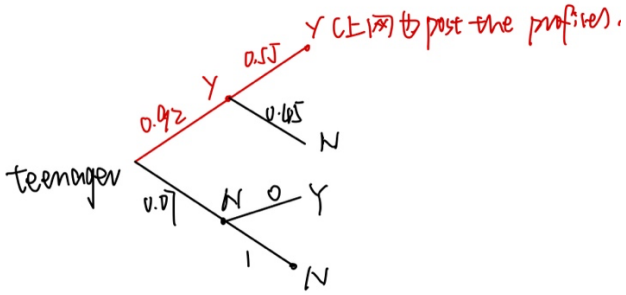
Through the tree diagram, the probabilities in the context are represented.

Example: Toss a coin



$$P(two\ heads) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

For many related problems, we can use both tree diagrams or general multiplication rule to solve them.
Sample question:
93% of teenagers (ages 12 to 17) use the internet; 55% of online teen have posted a profile. Q: inline and post the profile?
[Method 1] Tree diagram method

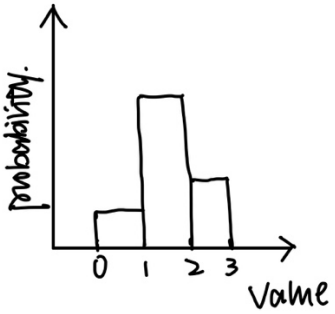


According to the tree diagram, $0.93 \times 0.55 = 51.15\%$
[Method 2] General multiplication rule

$$P(O \cap P) = P(O) \times P(P|O) = 0.93 \times 0.55 = 51.15\%$$

d. <u>Independence and mutually exclusive</u>
Independence $P(B|A) = P(B)$
Mutually exclusive $P(A\ and\ B) = 0$
独立不互斥 互斥不独立 两个不可能同时发生

5. Discrete random variables and mean (expected value) and SD
a. <u>Random variable</u>
A random variable takes <u>numerical values</u> that describe the outcomes of some <u>chance process</u>.



Probability distribution
The probability distribution of a random variable gives its possible values and their probabilities.
横轴：x；纵轴：P(x)

b. <u>Discrete and continuous</u>

Discrete	Continuous
x 是 numbers 是整数 不连续	x 是连续的整数 无法 list all values
e.g., 扔硬币 投篮	e.g., 学生高度 物品价格
$P(x \geq 7)$和$P(x > 7)$不同	单个值发生概率始终为 0 $P(x \geq 7) = P(x > 7)$

c. Mean (expected value)

$$\mu_x = E(x) = x_1p_1 + x_2p_2 + x_3p_3 + \cdots + x_ip_i = \sum x_ip_i$$

Interpretation: If many many XXX are randomly selected, the average number of XX is expected to be XX.

d. Standard deviation (and variance) of a discrete random variable

Variable

$$var(x) = \sigma_x^2 = (x_1 - \mu_x)^2p_1 + (x_2 - \mu_x)^2p_2 + (x_3 - \mu_x)^2p_3 + \cdots = \sum (x_i - \mu_x)^2p_i$$

Standard deviation

$$\sigma_x = \sqrt{\sum (x_i - \mu_x)^2\, p_i}$$

Interpretation: A randomly selected XXX will typically differ from the mean by about XX units.

6. 离散随机变量模型 1——二项分布 Binomial distribution

a. Binomial setting (condition)

1) Binary：只考虑成功或不成功
2) Independent：trials 之间相互独立，p 不变（每一次 trials 的结果无法 predict 下一次）
3) Number of trials：n 已知（重复该随机过程 n 次）
4) Success 的 p 已知且不变

b. 表示与含义

表示：B (n=_____; p=_____) 注意 n: number of trial 不是 repeat 的次数

含义：x=number of success when repeating n times（成功的概率为 p，不成功的概率为 1-p）

c. Binomial probability

计算器

$$p(x \geq 3) = Bcd\ (lower = \_\_\_\_\ upper = \_\_\_\_\ n = \_\_\_\_\ p = \_\_\_\_)$$

$$p(x = 3) = Bpd\ (x = \_\_\_\_\ n = \_\_\_\_\ p = \_\_\_\_)$$

1) $p(x = 3)$ 重复 n 次成功 3 次
2) $p(x \geq 3)$ 重复 n 次至少成功 3 次
3) $p(x > 3) = p(x \geq 4)$ 重复 n 次成功多于 3 次（计算器只能算≥和=）

公式

$$p(x = k) = \binom{n}{k}(p)^k(1-p)^{n-k}$$

p 为成功的概率；1-p 为不成功的概率

$\binom{n}{k}$又被写为 nCk，代表 n 个中选 k 个；例如$\binom{5}{3}$代表 5 个中选 3 个

d. Mean and standard deviation of a binomial distribution

1) $mean\ (\mu_x) = np$
2) $SD(\sigma_x) = \sqrt{np(1-p)}$
3) Shape: Approximately normal 条件$np \geq 10\ and\ n(1-p) \geq 10$（成次败次都出现 10 次）
4) Interpretation:
    a) Mean: on average, the expected number of success when repeating B for n times with a constant

succussing probability p is about to be μ.

b) SD: the typical difference between number of success and its mean when repeating B for n times with a constant probability p is about σ_x.

7. 离散随机变量模型 2——几何分布 Geometric distribution

a. Geometric setting (condition)

1) Binary：只考虑成功 p=成功概率
2) Independent：trials 之间相互独立，p is fixed

b. 表示与含义

G=number of trials…until success（总数不定 无 max，成功是在最后一次）

c. Geometric probability

某次成功（公式或计算器 Gpd）

$$p(G = x) = (1-p)^{x-1} \cdot p$$

$(1-p)^{x-1}$为失败的概率；p是成功的概率（最后一次）

累计成功（计算器 Gcd）

d. Mean and standard deviation of a geometric distribution

1) $mean = \frac{1}{p}$
2) $\sigma^x = \frac{\sqrt{1-p}}{p}$
3) Interpretation
   a) SD: values of _____ typically vary from μ by about σ, on average.
   b) Mean: in repeated sampling from the distribution of ____, the average of values will approach mean.

8. 随机变量的线性转换 Linear transformation of a random variable

Adding (or subtracting) the constant --- shape 不变 mean, median, quar $\pm a$

Multiplying (or dividing) by a constant --- spread, mean, med $\times b$

9. 独立随机变量的加减 Combing random variable

RV 线性转化：$\mu_{a+b} = a + b(\mu_x)$ $\sigma_{a+bx} = |b|(\sigma_x)$

RV 相加（满足 x 和 y 独立）：$\mu_{x+y} = \mu_x + \mu_{yi}$ $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$

RV 相减（满足 x 和 y 独立）：$\mu_{x-y} = \mu_x - \mu_{yi}$ $\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2$

Unit 5 Sampling Distributions (7-12%)

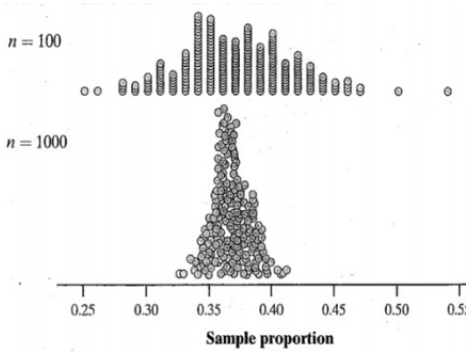
1. Sampling distribution 的建立

a. <u>Parameter and statistics</u>

	Parameter (population 值)	Statistics/estimator (sample 值)
Mean	μ	$\bar{x}$
Variance	σ^2	$s^2\ or\ s_x^2$
SD	σ	$s\ or\ s_x$
proportion	p	$\hat{p}$

b. <u>Unbiased estimator (center) and sampling variability (spread)</u>

1) Center
   a) Center=biased and unbiased estimator（center 与 n 无关，只和调查方法\公式有关）
   b) Biased estimator $var = \frac{1}{n}\sum(x_i - \bar{x})^2$ Unbiased estimator $var = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
   c) 对于一个无偏估计，Statistics' sampling distribution 的 mean 就是 parameter
2) Spread: sampling variability（sample mean 之间的差异）
   a) Low variability is better
   b) The larger n, the smaller sampling variability



2. Sampling distribution for proportion

	Sampling distribution for proportion	Sampling distribution of $\hat{p}_A - \hat{p}_B$
Mean	$\mu_{\hat{p}} = p$	$\mu_{\hat{p}_A - \hat{p}_B} = p_A - p_B$
Spread	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ $\sigma_{\hat{p}}\ max: when\ p\ approach\ 0.5\ and\ n \uparrow$	$\sigma_{\hat{p}_A - \hat{p}_B} = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$
Condition	**10% condition:** $n \le \frac{1}{10}N$ **Large count:** $np \ge 10\ \ n(1-p) \ge 10$	**10% condition:** $n_1 \le \frac{1}{10}N_1\ \ n_2 \le \frac{1}{10}N_2$ **2 independent random sample** (Dependent: 2 proportions in 1 sample) **Large count:** $np \ge 10\ \ n(1-p) \ge 10$
If not dependent	SD 变小（因为剪掉了斜方差）	

3. Sampling distribution of sample mean

	Sampling distribution of sample mean	Sampling distribution of difference in sample mean
Mean	$\mu_{\bar{x}} = \mu$	$\mu_{\bar{x}_A - \bar{x}_B} = \mu_A - \mu_B$

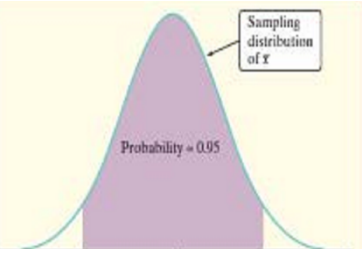
	Sampling distribution of sample mean	Sampling distribution of difference in sample mean
Spread	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ The larger n, the lower variability	$\sigma_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$
Condition	**Normal** 1) 原 population 是 normal，则直接定义为 normal 2) Central limit theorem：原 population 非 normal，但满足 $x \geq 30$（定义为 normal） **10% condition when sampling without replacement** N is at least 10 (n) --- independent	**Normal** 1) 原 population 是 normal，则直接定义为 normal 2) Central limit theorem：原 population 非 normal，但满足 $x \geq 30$（定义为 normal） **10% condition when sampling without replacement** N_A is at least 10 (n_A) --- independent N_B is at least 10 (n_B) --- independent **2 independent random sample** (Dependent: 2 proportions in 1 sample)

Prep Session for Inference Statistics (Confidence Interval Basics and Significance Test Logics)

1. Confidence interval basics

a. <u>The purpose and idea of confidence interval</u>

1) 置信区间的目的：estimate parameter
2) Point estimator：估计 parameter 的一个 statistic；Point estimate：在 statistic 的一个具体值
3) The idea of confidence interval
   a) 在我们重复取样的过程中，有 95%的样本平均值会落在区间（红色区域）内
   b) 这是我们对 μ 的一个 95%的置信区间
   c) Population 值是否在 interval 中

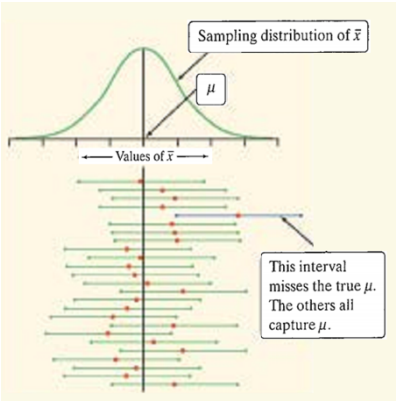


b. <u>The definition of confidence interval, margin of error and confidence level</u>

1) 置信区间 confidence interval：A C% confidence interval gives an interval of plausible values for a parameter. The interval is calculated from the data and has the form:

$$point\ estimate \pm margin\ of\ error\ (ME)$$

2) 边际误差 margin of error：The difference between the point estimate and the true parameter value will be less than the margin of error in C% of all sample.
3) 置信水平 confidence level C%：The confidence level C gives the overall success rate of the method for calculating the confidence interval. That is, in C% of all possible samples, the method would yield an interval that captures the true parameter value.



$$95\% = \frac{\text{包含 parameter 的 CI 们}}{\text{重复 100 个 samples 得到的 CI}}$$

例如：重复取样 200 次，有 190 次 sample 得到的 interval 抓住了真实的 parameter，有 10 次没抓住。Confidence level=190/200=95%。

4) Interpreting the confidence interval: We are C% confident that the interval from ____ to ____ captures the [parameter in context].

2. Significance test — logics

a. <u>The definition of the significance test</u>

1) Significance test：a formal procedure for comparing observed data with a claim (also called a hypothesis), the truth of which is being assessed.
2) 根据 sample 对比 population 看它是不是有变化（有显著差异）

b. <u>State hypothesis</u>

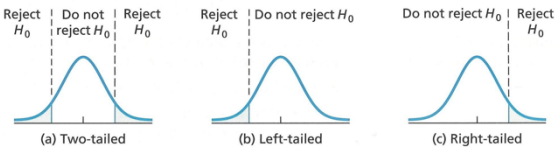
1) 原假设 null hypothesis：H_0；备择假设 alternative hypothesis：H_a（两者针锋相对）
2) H_0：population 值 = 错误前提

3) Ha

a) population 值 > 错误前提 (right-side test)

b) population 值 < 错误前提 (left-side test)

c) population 值 ≠ 错误前提 (two-side test)



	Two-tailed test	Left-tailed test	Right-tailed test
Sign in H_a	≠	<	>
Rejection region	Both sides	Left side	Right side

c. <u>Conclusion</u>

P-value < α. We reject H_0 and we have convincing evidence that H_a is correct.

P-value > α. We fail to reject H_0 and we do not have convincing evidence that H_a is correct.

3. Power of the test and two types of error

a. Type I 和 Type II 结论

	Type I 结论		Type II 结论	
Decision	P-value < α Reject H_0 and H_a is correct		P-value > α Fail to reject H_0 and no evidence that H_a is correct	
Truth	与 I 结论一致	与 I 结论相反	与 II 结论一致	与 II 结论相反
	H_a is correct H_0 is incorrect	H_a is incorrect H_0 is correct	H_a is incorrect H_0 is correct	H_a is correct H_0 is incorrect
	Power of the test	Type I error	刚好重做	Type II error

b. <u>Power of the test and two types of error</u>

$$\text{power of the test} = P(\text{得出 Type I 结论} | H_a \text{ is correct})$$

$$\text{type I error} = P(\text{得出 Type I 结论} | H_0 \text{ is correct}) = \text{the defi of } \alpha$$

$$\text{type II error} = P(\text{得出 Type II 结论} | H_a \text{ is correct}) = \beta$$

The relationship between two types of error and power of the test

$$\text{power of the test} = 1 - \beta \text{ (type II error)}$$

$$\text{type I error} \uparrow \quad \text{power of the test} \uparrow \quad \text{type II error} \downarrow$$

c. <u>How to increase the power of test</u>

1) Increasing sample size (↑ n)
2) Increasing the significance level (↑ α)
3) Increasing the difference between the null and alternative
4) Parameter values that is important to detect（选一个更远的 parameter）

Unit 6 and Unit 8 Inference for Categorical Data — Proportion and Chi-Square (12-15% for Unit 6; 2-5% for Unit 8)

1. Inference for categorical data about proportions

a. <u>One-proportion-z-interval for population proportion</u>

1) Condition
   a) 10% condition: when sampling without replacement $n \leq \frac{1}{10}N$
   b) Large count condition: $n\hat{p} \geq 10 \; and \; n(1-\hat{p}) \geq 10$
   c) Random sample: well-designed random sample
2) Do
   a) Using 1-prop-z-interval with C%=_____ x=_____ n=_____, we find the interval (_____, _____).
   b) $\hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (point estimate: $\hat{p}$; standard error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; critical value: Z^*)
   c) $\hat{p}$越接近 1/2，ME 越大
3) Conclude: we are c% confident that this interval from _____ to _____ captures the true proportion of _____.
4) FRQ 问能否使用 confidence interval 来 determine something 时，先回答：all values within this interval are plausible values for the true proportion.

b. <u>Two-proportion-z-interval for population proportions</u>

1) Condition
   a) Random: two independent random sample or two groups in the randomized experiment
   b) 10% condition: $n_1 \leq \frac{1}{10}N_1 \;\; n_2 \leq \frac{1}{10}N_2$ (experiment do not need to check the 10% condition)
   c) Large count condition: $n_1\hat{p}_1 \geq 10 \; and \; n_1(1-\hat{p}_1) \geq 10 \quad n_2\hat{p}_2 \geq 10 \; and \; n_2(1-\hat{p}_2) \geq 10$
2) Do
   a) Using 2-prop-z-interval with C%=_____ x_1=_____ n_1=_____ x_2=_____ n_2=_____, we find the interval (_____, _____).
   b) $\hat{p}_A - \hat{p}_B \pm Z^* \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{n}}$ (point estimate: $\hat{p}_A - \hat{p}_B$; standard error: $\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{n}}$; critical value: Z^*)
3) Conclude: we are c% confident that this interval from _____ to _____ captures the true difference between _____ to _____ (P_A-P_B).
4) 若问能否用 confidence interval 来 determine something。若 interval 又负、0、正，则无法判断 (no exactly difference)

c. <u>One-proportion-z-test for population proportion</u>

1) State
   a) $H_0: p = 80\% \; H_a: p > 80\%$ (right-tail test)
   b) $H_0: p = 80\% \; H_a: p < 80\%$ (left-tail test)
   c) $H_0: p = 80\% \; H_a: p \neq 80\%$ (two-tail test)
2) Condition

a) 10% condition: when sampling without replacement $n \le \frac{1}{10}N$

b) Large count condition: $n\hat{p} \ge 10 \; and \; n(1-\hat{p}) \ge 10$

c) Random sample: well-designed random sample

3) Do

a) Using 1-prop-z-test, with x=_____ n=_____, we find z=_____ and p-value=_____.

b) Test statistic: $Z_{\hat{p}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

4) Conclude

a) P-value < α. We reject H_0 and we have convincing evidence that H_a is correct.

b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that H_a is correct.

d. <u>Two-proportion-z-test for population proportions</u>

1) State

a) $H_0: p_A = p_B \;\; p_A - p_B = 0 \qquad H_a: p_A > p_B \;\; p_A - p_B > 0$ (right-tail test)

b) $H_0: p_A = p_B \;\; p_A - p_B = 0 \qquad H_a: p_A < p_B \;\; p_A - p_B < 0$ (left-tail test)

c) $H_0: p_A = p_B \;\; p_A - p_B = 0 \qquad H_a: p_A \ne p_B \;\; p_A - p_B \ne 0$ (two-tail test)

2) Condition

a) Random: two independent random sample or two groups in the randomized experiment

b) 10% condition: $n_1 \le \frac{1}{10}N_1 \;\; n_2 \le \frac{1}{10}N_2$ (experiment do not need to check the 10% condition)

c) Large count condition: $n_1\hat{p}_1 \ge 10 \; and \; n_1(1-\hat{p}_1) \ge 10 \quad n_2\hat{p}_2 \ge 10 \; and \; n_2(1-\hat{p}_2) \ge 10$

3) Do

a) Using 2-prop-z-test, with x_A=_____ n_A=_____ x_B=_____ n_B=_____, we find z=_____ and p-value=_____.

b) Test statistic: $Z_{\hat{p}_A - \hat{p}_B} = (\hat{p}_A - \hat{p}_B) - 0 \Big/ \sqrt{\hat{p}_c(1-\hat{p}_c)(\frac{1}{n_A}+\frac{1}{n_B})} \qquad \hat{p}_c = \frac{x_A + x_B}{n_A + n_B}$

4) Conclude

a) P-value < α. We reject H_0 and we have convincing evidence that H_a is correct.

b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that H_a is correct.

2. Inference for categorical data: Chi-square

a. <u>The chi-square statistic χ^2</u>

1) 卡方的目的：Comparing observed and expected counts

2) The chi-square statistic is a measure of how far the observed counts are from the expected counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

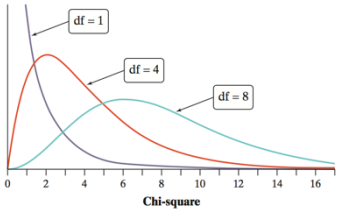
where the sum is over all possible values of the categorical variable.

3) χ^2 distribution

a) χ^2没有<0 的 negative value

b) χ^2 distribution 由 df 决定；其 shape 随着 df 的改变而改变

c) df 越大，skewness 越小



4) The χ^2 distribution and p-value (take an example)

a) Example：公司宣称牛奶巧克力糖将含有棕色和红色各 13%，黄色 14%，绿色 16%，橙色 20%，蓝色 24%

b) do a random sample of 60 candies and record the results as

Color	Observed
Blue	9
Orange	8
Green	12
Yellow	15
Red	10
Brown	6

c) calculate expected counts $expected\ counts = np$

Orange: $(60)(0.20) = 12.00$
Green: $(60)(0.16) = 9.60$
Yellow: $(60)(0.14) = 8.40$
Red: $(60)(0.13) = 7.80$
Brown: $(60)(0.13) = 7.80$

Color	Observed	Expected
Blue	9	14.40
Orange	8	12.00
Green	12	9.60
Yellow	15	8.40
Red	10	7.80
Brown	6	7.80

d) calculate χ^2 and df $(df = k - 1 = 6 - 1 = 5)$

$$\chi^2 = \frac{(9-14.40)^2}{14.40} + \frac{(8-12.00)^2}{12.00} + \frac{(12-9.60)^2}{9.60} + \frac{(15-8.40)^2}{8.40} + \frac{(10-7.80)^2}{7.80} + \frac{(6-7.80)^2}{7.80}$$
$$= 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 = 10.180$$

e) get the p-value according to the table c

Table C Chi–square distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
→ 5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02

b. χ^2-GOF-test (chi-square-test for goodness of fit)

1) State

a) H_0: the company's stated XXX distribution for XXX is correct.

b) H_a: the company's stated XXX distribution for XXX is not correct.

c) Or: H_0: $p_{blue} = 0.24$ $p_{orange} = 0.20$ $p_{green} = 0.16$ ...

Ha: at least two of p_i's are incorrect.

2) Condition（要计算 expected value 并逐一写在 condition 部分）

a) Random: random sample

b) 10% condition: when sampling without replacement $n \le \frac{1}{10}N$

c) Large count condition: according to my calculation, all expected counts are at least 5.

3) Do（使用计算器时，expected value 放在 list 2，且需要手算）

a) Using χ^2-GOF-test with df=____, we find that χ^2=_____ and p-value=_____.

4) Conclude

a) P-value < α. We reject H_0 and we have convincing evidence that the company's stated XXX distribution is not correct.

b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that the company's stated XXX distribution is not correct.

5) Follow-up test for the χ^2-test

a) CNTRB value: detail contribution

b) 计算器中得到的 CNTRB 会出现在 List 3 中，CNTRB 越大说明单个值对总的卡方检验贡献越多

c) $CNTRB = {(O-E)^2}/{E}$

d) 因此，CNTRB 最大的那个变量，我们称其：XXX is largest contribution to the χ^2 statistic.

c. χ^2-test for Homogeneity

1) Purpose of χ^2-test for Homogeneity: comparing distribution of a categorical variable（数据来自两个或两个以上的独立随机样本，以判断几次重复试验结果是否一致）

2) State

a) H0: there is no difference in the true distributions of XXX.（三种音乐下的 entrée 分布相同）

b) Ha: there is a difference in the true distributions of XXX.（三种音乐下的 entrée 分布不同）

3) Condition

a) Random: several independent random samples or groups in randomized experiments were used

b) 10% condition: $n \le \frac{1}{10}N$

c) Large count condition: according to my calculation, all expected counts are at least 5.

d) Expected counts table（计算器做完 test 后自动把 expected matrix 生成在 Mat B 中，在条件部分，此表格必须抄上）

$$\text{expected counts} = \frac{(\text{行 total})(\text{列 total})}{\text{总 total}}$$

4) Do（在计算器使用时，只需要在 observed 里输入 Mat A 即可，不需要在 expected 输入 Mat B）

a) Using technology χ^2-test for Homogeneity with df=___, we find χ^2= ___ and p-value=____

5) Conclude

a) P-value < α. We reject H_0 and we have convincing evidence that there is a difference in the true distribution of [entree ordered] when [no music, French music, or Italian music is played].

b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that there is a difference in the true distribution of XXX when XXX.

6) Follow-up test for the χ^2-test

a) SPSS\MATLAB 等软件输出结果才有 CNTRB

Chi-Square Test: None, French, Italian

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

		None	French	Italian	Total
1	Obs	30	39	30	99
	Exp	34.22	30.56	34.22	
	CNTRB	0.521	2.334	0.521	
2		11	1	19	31
		10.72	9.57	10.72	
		0.008	7.672	6.404	
3		43	35	35	113
		39.06	34.88	39.06	
		0.397	0.000	0.422	
Total		84	75	84	243

Chi-Sq = 18.279, DF = 4, P-Value = 0.001

d. χ^2-test for Independence

1) Chi-sq test for independence and homogeneity 区别
   a) Independence: one single random sample
   b) Homogeneity: several independent random samples or groups in a randomized experiment.
   c) Purpose of χ^2-test for Independence: find the evidence of an association between two categorical variables from a single sample.（对于一个 sample 的两个不同特征，在统计学上证明有无关联）
   d) 相同点：Two-way table，Do 完全相同
2) State
   a) H_0: there is no association between X and Y (independent).
   b) H_a: there is an association between X and Y (no independent).
3) Condition
   a) Random: a random sample
   b) 10% condition: $n \leq \frac{1}{10}N$
   c) Large count condition: according to my calculation, all expected counts are at least 5.
4) Do（在计算器使用时，只需要在 observed 里输入 Mat A 即可，不需要在 expected 输入 Mat B）
   a) Using technology χ^2-test for Homogeneity with df=___, we find χ^2= ___ and p-value=____
5) Conclude
   a) P-value < α. We reject H_0 and we have convincing evidence that there is an association between X and Y.
   b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that there is an association between X and Y.

e. 卡方注意事项

1) Total 不算表格中的列或行，不输入；
2) 是 expected 大于等于 5！不是 obs；
3) 要手算 expected counts 并列表；
4) Chi-square test 输入的都是 counts 不是 proportion；
5) Chi-square test 检测 obs 和 exp 是否有差异；
6) Follow-up analysis 检测谁和 exp 差异最大；
7) 学会区分 two way 的两种 chi-square test。

Unit 7 and Unit 9 Inference for Quantitative Data — Means and Slopes (10-18% for Unit 7; 2-5% for Unit 9)

1. t-distribution

因为不知道 population SD，使用 t-distribution 代替 z-distribution。相比于 z-distribution，对于 t-distribution 有一些特点

- $\mu = 0$相同，同样 symmetric and unimodal，公式与 z 一样
- 比 normal 更佳 spread out（df 与 n 有关）
- $df \downarrow$ tail 两边面积越大
- $df \uparrow$ 越接近 normal

2. Inference for quantitative data about means

a. <u>One-sample-t-interval for population mean</u>

1) Condition
   a) Random: a random sample or a randomized experiment
   b) 10% condition: $n \leq \frac{1}{10}N$
   c) Normal（满足其一）
      i. Central limit theorem: $n =$ _____ ≥ 30
      ii. 原 population is normal
      iii. $n < 30$, check strong skewness or obvious outlier through dotplot, boxplot, stemplot or histogram (if no—assume population is normal)
2) Do
   a) Using 1-sample t-interval, $\bar{x} =$ _____ $S_x =$ _____ $C\% =$ _____ $n =$ _____, we find the interval (_____, _____), df=_____(n-1).
   b) t-interval for mean: $\bar{x} \pm t^* \frac{S_x}{\sqrt{n}}$ $(when\ \sigma\ is\ unknown)$
3) Conclude: we are c% confident that this interval from _____ to _____ captures the true mean of _____.
4) Width of ME
   a) C% increase ME increase
   b) N increase ME decrease
   c) Sample SD increase ME increase
   d) ME is NOT related to center ($\bar{x}$)

b. <u>Two-sample-t-interval for difference between means ($\mu_A - \mu_B$)</u>

1) Condition
   a) Two independent random samples or two groups in a randomized group
   b) 10% condition: $n_1 \leq \frac{1}{10}N_1$ $n_2 \leq \frac{1}{10}N_2$ (experiment do not need to check the 10% condition)
   c) Normal（n_1和n_2各自满足其一）
      i. Central limit theorem: $n =$ _____ ≥ 30
      ii. 原 population is normal
      iii. $n < 30$, check strong skewness or obvious outlier through dotplot, boxplot, stemplot or

histogram (if no—assume population is normal)

2) Do
   a) Using 2-sample t-interval for $\mu_A - \mu_B$, $C\% =$ _____ $\bar{x}_A =$ _____ $S_A =$ _____ $n_A =$ _____ $\bar{x}_B =$ _____ $S_B =$ _____ $n_B =$ _____ df=_____, we find the interval (_____, _____).
   b) t-interval for mean: $(\bar{x}_A - \bar{x}_B) \pm t^* \sqrt{\frac{{S_A}^2}{n_A} + \frac{{S_B}^2}{n_B}}$ $(when\ \sigma\ is\ unknown)$
   c) df is calculated by calculator.
3) Conclude: we are c% confident that this interval from _____ to _____ captures the true difference of the means (X-Y) of _____.

c. <u>One-sample-t-test for population mean</u>

1) State
   a) H0: $\mu = 100$
   b) Ha: $\mu > 100\ (right - tail);\ \mu < 100\ (left - tail);\ \mu \neq 100\ (two - tail)$
2) Condition
   a) Random sample
   b) 10% condition: $n \leq \frac{1}{10}N$
   c) Normal（满足其一）
      i. Central limit theorem: $n =$ _____ ≥ 30
      ii. 原 population is normal
      iii. $n < 30$, check strong skewness or obvious outlier through dotplot, boxplot, stemplot or histogram (if no—assume population is normal)
3) Do
   a) Using 1-sample-t-test for μ, with $\bar{x} =$ _____ $S_x =$ _____ $df(n-1) =$ _____ $n =$ _____, we find $t =$ _____ and $p\ value =$ _____.
   b) $t = {\bar{x} - \mu_0}\Big/{\frac{S_x}{\sqrt{n}}}$
   c) $df = n - 1$
4) Conclude
   a) P-value < α. We reject H_0 and we have convincing evidence that true mean of ____ is correct.
   b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that true mean of ____ is correct.

d. <u>Matched-paired t-interval for population mean</u>

1) 有 difference 不用 sample mean，每个 difference 作为一个新的 sample
2) Condition
   a) Random: randomized experiment or random sample
   b) 10% condition: $n \leq \frac{1}{10}N$
   c) Normal（满足其一）
      i. Central limit theorem: $n =$ _____ ≥ 30
      ii. 原 population is normal
      iii. $n < 30$, check strong skewness or obvious outlier through dotplot, boxplot, stemplot or

histogram (if no—assume population is normal)

3) Do
   a) Using paired-t-interval with $\bar{d} =$ ____ $S_d =$ ____ $n =$ ____ $C\% =$ ____, we find (____, ____)
   b) $\bar{d} \pm t^* {S_d}/{\sqrt{n}}$ $(df = n_1 - 1)$
4) Conclude: we are C% confident that this interval from _____ to _____ captures the true mean difference of _____ between _____ and _____ (X minus Y).

e. <u>Matched-paired t-test for mean difference</u>

1) State
   a) H0: $\mu_D = 0$
   b) Ha: $\mu_D > 0$ $\mu_D < 0$ $\mu_D \neq 0$ (difference=X-Y)
2) Condition
   a) Random: randomized experiment or random sample
   b) 10% condition: $n \leq \frac{1}{10}N$
   c) Normal（满足其一）
      i. Central limit theorem: $n =$ _____ ≥ 30
      ii. 原 population is normal
      i. $n < 30$, check strong skewness or obvious outlier through dotplot, boxplot, stemplot or histogram (if no—assume population is normal)
3) Do
   a) Using paired-t-test with $\bar{d} =$ ____ $\mu_d =$ ____ $n =$ ____ $S_d =$ ____, we find that t=____ p-value=____.
   b) $t = {\bar{d} - \mu_0}/{\frac{S_d}{\sqrt{n}}} = {\bar{d} - 0}/{\frac{S_d}{\sqrt{n}}}$ $df = n - 1$
4) Conclude
   a) P-value < α. We reject H_0 and we have convincing evidence that XXX.
   b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that XXX.

f. <u>Two-sample-t-test for difference between means ($\mu_A - \mu_B$)</u>

1) State
   a) H0: $\mu_A - \mu_B = 0$
   b) Ha: $\mu_A - \mu_B > 0$ $\mu_A - \mu_B < 0$ $\mu_A - \mu_B \neq 0$
2) Condition
   a) Two independent random samples or two groups in a randomized group
   b) 10% condition: $n_1 \leq \frac{1}{10}N_1$ $n_2 \leq \frac{1}{10}N_2$ (experiment do not need to check the 10% condition)
   c) Normal（n_1和n_2各自满足其一）
      i. Central limit theorem: $n =$ _____ ≥ 30
      ii. 原 population is normal
      iii. $n < 30$, check strong skewness or obvious outlier through dotplot, boxplot, stemplot or histogram (if no—assume population is normal)
3) Do

a) Using 2-sample t-test for $\mu_A - \mu_B$ with $\bar{x}_A =$ _____ $S_A =$ _____ $n_A =$ _____ $\bar{x}_B =$ _____ $S_B =$ _____ $n_B =$ _____ df=_____, we find that t=____ and p-value=_____.

b) $t = (\bar{x}_A - \bar{x}_B) - 0 \Big/ \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$

4) Conclude
   a) P-value < α. We reject H_0 and we have convincing evidence that the true difference of the means (A-B) of _____.
   b) P-value > α. We fail to reject H_0 and we do not have convincing evidence that the true difference of the means (A-B) of _____.

3. Inference for quantitative data about slope

a. <u>Sample slope and true slope</u>

Sample slope: b

True slope (population slope): β

b. <u>统计学软件解读</u>

	Minitab			
Predictor	Coef.	SE coef.	T	P
Constant	38257 (y-intercept)	2446	15.64	0
Miles driven (x)	-0.1629 (slope)	0.03096 (SE_b)	-5.26 (双尾的 t)	0

c. <u>Linear regressions t-interval for slope (β)</u>

1) Condition
   a) Linear: the actual relationship between x and y is linear（用 x 和 y 的 scatterplot 或 residual plot 进行判断）
   b) Equal SD: the SD of y is the same for an values of x（不能越来越离散）
   c) Normal
      i. Check dotplot of residual (1-var)---no skewness or outliers
      ii. If skewed or outliers, $n \geq 30$
   d) Random: random sample or randomized experiment
   e) Independence: $n \leq \frac{1}{10}N$
2) Do: $b \pm t^* \cdot SE_b$ $(df = n - 2)$
3) Intercept the standard error (SE): if we repeated the random assignment many times, the slope of the sample regression line would typically vary by about _____ from the slope of the true regression line for the _____ (predicted y) from _____ (x)

d. <u>Linear regressions t-test for slope (β)</u>

1) State
   a) H0: $\beta = 0$ （x 和 y 无线性相关）
   b) Ha: $\beta \neq 0$ （有线性相关） $\beta > 0$ （正相关） $\beta < 0$ （负相关）
2) Condition
   a) Linear: the actual relationship between x and y is linear（用 x 和 y 的 scatterplot 或 residual plot

进行判断）

b) Equal SD: the SD of y is the same for an values of x（不能越来越离散）

c) Normal

i. Check dotplot of residual (1-var)---no skewness or outliers

ii. If skewed or outliers, $n \geq 30$

d) Random: random sample or randomized experiment

e) Independence: $n \leq \frac{1}{10}N$

3) Do: $t = \frac{b-\beta}{SE_b} = \frac{b-0}{SE_b} = \frac{b}{SE_b}$

4. Transforming to achieve linearity

把一个不 linear 的 model 通过数学变换，变得 linear（例如给 all data ×2 或平方）