Statistics

2022-07-17

1 第一篇数据产生以及统计描述

1.1 第一章描述性统计学探索数据

1. 图形描述
2. 数字描述

- 数据集中的描述：均值与中位数，数据对称时均值与中位相等，均值向数据分布（右偏或左偏）偏离。
- 数据分散的描述：interquartile range = third quartile - 1st quartile。标准差是用于描述数据分散的常用数值。
- 均值和标准差都对少部分很大或很小的值很敏感,可以考虑使用中位数和 interquartile range

1.2 第二章数据产生，抽样

总体（population): entire group we want information
参数（parameter)：quantity about the population we are interested in. 样本（sample): part of population from which we collect information.
统计量（estimate，statistic)：the quantity we are interested in as measured in the sample.

关键：即使一个很小的样本也能产生一个于总体参数相近的估计。

抽样方法：

1. 不放回的简单随机抽样（a simple random sampling)。
2. 分层的随机抽样 (a stratified random sample)：相似的为一层，然后，在每一层进行一个简单的随机抽样，然后把这些样本组合起来。

Bias and chance error:

- Bias(systematic error)

1. selection bias：a sample of convenience make it more likely to sample certain subjects than others

2. non-response bias: less likely to answer a question at a special situation

3. voluntary response bias: websites that post reviews of businenss are more likely to get response from customer who had very bad or very good experiences.

- Chance error(sampling error): 随机抽样, 估计和总体的偏差, 每一次的抽样有不同的 chance error。抽样的随机性。

$$estimate = parameter + bias(systematic\ error) + chance\ error(sampling\ error)$$

Note: 增大样本可以减少 chance error，并且我们可以计算 chance error 具体有多大。但是增大样本只是让 bias 在一个更大的规模上重复，并且我们不能知道 bias 的大小。

因果分析与关联分析

observation study: 测量一个感兴趣的事的结果，用于观察关联关系。

Association is not causation, there may be confounding factors

Causation experiment(randomized controlled experiments): 1. Subjects are assigned into treatment and control groups. at random.
2. Subjects in controlled group get a placebo to ensure both groups equally affected by the placebo effect: the didea of being treated may have an effect by itself.
3. Double-blind.

More：
The wired power of placebo effect, explained

2 第二篇当理想照进现实——从总体分布到样本分布

2.1 第一章大数定理与中心极限定理

Three histograms: 1. Probability histogram for producing the data. 2. The histogram of 100 observed tosses. 3. The probability of the statistic.

Law of large numbers: When sample size is large enough, the $\bar{x}_N$ will be likely close to μ . 1) applied for averages and percentages, but not for sums. 2) sampling with replacement from a population or for simulating data from a probability histogram.

More advanced large number laws: the empirical histaogram will be close to probability to histogram producing the data.

Central limit theory: the sample sum statistic(averages and percentages are sums in disguise) distribution is normal distribution

应用条件：放回抽样，或者每次都从同一个概率分布函数抽样（其实不同的也可以?）
Sample size is large enough.(if no strong skewness, n > 15 is sufficient)

2.2 第二章概率

Standard definition: proportion of times this event occurs in many repetitions.
Subjective probability: not based on experiments, different people assign different subjective probabilities to the same event.

Four basic rules

Complement rule: P(A does not occur) = 1 - P(A)

Rules for equally likely outcomes: $P(A) = \frac{number\ of\ outcomes\ in\ A}{n}$

Addition rule: A and B are mutually exclusive(don't occure at the same time), then:

$$P(AorB) = P(A) + P(B)$$

Multiplication rule: A and B are independent(one occures doesn't change the probability that the other occurs),then:

$$P(AandB) = P(A)(B)$$

条件概率（conditional probability)

$$P(B|A) = \frac{P(AandB)}{P(A)}$$

General multiplication rule: $P(AandB) = P(A)P(B|A)$, special case where A and B are independent: $P(AandB) = P(A)P(B)$.

Bayes's rule Bayesian analysis

False positives case warner's randomized response model

2.3 第三章正态分布与二项分布

Normal curve: bell-shaped.
Empirical rule:
About 2/3(68%) fall within one sd of the mean.
About 95% fall within 2 sd of the mean.
About 99.7 fall within 3 sd of the mean.

Standardize data: $z = \frac{height - \bar{x}}{s}$
z meas how many sd the height away from the mean. no unites.

Normal approximation: 1. Finding areas under teh normal curve.(we can look up area to the left of a given value) the empirical rule is a special case of normal approximation.
2. Computing percentiles for normal data: 30% data for normal curve, the height is z sd away from mean.

Binomial probability

$$\frac{X(success\ count) - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

Note: 简单随机抽样是不放回的抽样，不是二项分布设定，因为每取出一个, 概率 P 就，改变了; 但是如果总体 size 远大于样本 size，那么放回抽样和不放回抽样就是大致一致的，服从于二项分布，服从于正态曲线。

2.4 第四章样本分布

Expected value of the sample average , E($\bar{x}_N$) is the population average. Standard error: statistic's sd(其实就是样本统计量的标准差), tells us roughly how far off the statistic will be from it expected value.

Expected value and *SE* for average > E($\bar{x}_N$) = μ (Square root law), $SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$

1. More lager sample size n, more smaller SE, it can be used to determine sample size to get desired accuarcy
2. SE don't depend on the size of the population, only on the size of the sample.

Expected value and *SE* for sum > $E(S_n) = n $, $SE(S_n) = \sqrt{n}\sigma$

Expected value and *SE* for percentages Framework for counting and classifying:

$$E(percentage\ of\ 1s) = 1\mu 100\frac{\sigma}{\sqrt{n}}100$$

Expected value and *SE* when simulating A random variable X that is simulated has K possible outcomes , $\mu = \sum_{i=1}^{k} x_i P(X = x_i), \sigma^2 = \sum i = 1^k (x_i - \mu)^2 P(X = x_i)$

3 第三篇一叶知秋——如何预知未来

3.1 第一章线性回归

Scatter plot three element: direction(slope up or down), form(points cluster around a line or other), strength(how closethe points follow the form)

Summary of pair data: $\bar{x}$, s_x, $\bar{y}$, s_y, r

How to quantify the strength?
If it is liner former, the correlation coefficient r is a good choice. standardized $x * y$,not affected by the scale of either variable. its sign gives the direction and its absolute value gives the strength.

Note: r is only useful for measuring linear association.and correlation does not mean causation

How to get the regression line?
To minimize the MSE(mean squared error), the method of least squares gives the analytic answer: $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$. This line $y = a + bx$ is called the regression line.

Another interpretation of the regression line:
> Computes the average value of y when the first coordinate is near x.
Note: The average often times is the best estimate when no extra information is provided.

向均值回归? regression effect(回归效应)?
因为 1）$\bar{x}$ 的预测值是 $\bar{y}$，2）$b = r\frac{s_y}{s_x}$ ，也就是说当 x 偏离 $\bar{x}$ 一个 sd 时，y 只向 $\bar{y}$ 偏离 $r * sd$ 个单位，也就是 y is fewer sd away from $\bar{y}$ than x is from $\bar{x}$.
i.e. Football shaped scatter, exam scores. my becaused by regression fallacy.

Note: x to y and y to x are two different regression line, cannot predict each other.

回归中的正态估计?
在回归线（football shape scatter) 中的某一点 x 处，y 服从于正态分布即：$\frac{Y-y(predict)|x}{\sqrt{1-r^2}s_y}$

如何检查回归使用是否正确?
Residual plot. 残差就真实值与预测值的差. it should be a unstructured horizontal band. curved plot: not liner;but the data can be $\surd$ or log transformation to liner to analyze.

scatter arises: heteroscedastic(may produce homoscedastic by y variable transformation, and it may result in a non-liner scatter , which require a second transformation in of x to fix)

离群值 outliers? 离 x 均值很远的 x 可能会对回归线的构建有很大的影响（influential point）, 会使得回归线向它偏离，无法用残差图检验。

一些问题：- 预测 y 时，x 应该在其范围之中，超出 x 的取值范围以后可能就不是线性关系。- 对总结数据注意，比如平均值，它们的变化更小，相关性?
R^2, 可以被回归线解释的部分，$1 - r^2$ 就是不能解释的，就是残差。

4 第四篇我们距离真相还有多远——置信区间和假设检验

4.1 第一章 Confidence interval

SE gives the chance error, confidence interval give a more precise statement.
已知一个样本统计量的分布，那么每一次的抽样我们可以说有 95% 几率该样本的统计量大小不会

偏离该总体参数的 2 个 SE(如果该统计量服从于正态分布)，也就是可以说每一次抽样得到到统计量，我们都有 95% 的把握说总体参数不会偏离超过 2SE 于该统计量。每一次的抽样都可以得到一个置信区间。
注意：置信区间随着每一次抽样变化而变化，但是总体参数是一个固定的值。

confidence = estimate +/- zSE(if statistic ~ normal distribution)

总体方差差未知?
bootstrap：用样本方差代替得到一个估计置信区间。

More：置信区间的大小由 zSE 决定，称作 *margin of error*, 因为 $SE = \frac{\sigma}{\sqrt{n}}$ ，所以可以通过增大样本 *size* 减少区间。同样也可减少 z 减少区间，比如 80% 区间。
百分数的 95% 置信区间：estimated percentage +/- $\sqrt{n}$
因为 $\sigma = \sqrt{p(1-p)} < \frac{1}{2}$

4.2 第二章假设检验原理

假设检验的逻辑?
设定零假设，备择假设，收集数据并评估该数据是否满足零假设从而接受零假设或拒绝，零假设一般是什么都没有发生，所以我们想要拒绝它，导致假设检验的逻辑不是很直接。

检验统计量? test statistic
A test statistic measures how far away the data are from what we would expect if H0 is true. i.e z-statistic:

$$z = \frac{observed - expected}{SE}$$

观测值是一个用于评估 H0 的统计量, expected and SE are the expected value and SE of the this statistic, computed under the assumption H0 is true.

p value is the probability of getting a value of z as extreme or more extreme than the observed z, assuming HO is true.

Note:Ho 是否正确是一个确定的事，p 值只是给出了在 H0 为真的假设下, 观察到如此极端值得概率大小。

实际上当我们进行 z 检验时，我们用观测统计量 - 期望统计量，然后除以统计量的 SE，计算出观测统计量在零假设的情况下偏离期望值多少个 *sd*。实际上样本的统计量的 SE 需要用总体方差进行计算，在 sample size > 20 的情况下可以直接用样本 *sd* 代替总体 *sd*，进行近似求解

sample SE。如果 sample size < 20，用样本 sd 代替总体 sd 计算那么其服从于 t(n-1), 置信区间：$\bar{x} + / - t_{n-1}SE$

其他：1. 统计学上的显著不能说明效应大小很重要，因为大的样本数目可以减少 SE, 使得样本统计量的分布更加集中，那么只要一个很小的偏离就可以具有统计学上的显著。
2. 95% 的置信区间包括了所有零假设不会被拒绝的值，对于一个双端检验 p 值为 0.05。
3. 两类错误：H0 为真, 拒绝了 type1 false positive,H0 为假，接受了 type2 error false negative

4.3 第三章假设检验之 z 检验以及 t 检验

Two sample z-test

$$z = \frac{observed\ difference - expected\ difference}{SE\ of\ difference} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{SE\ of\ difference}$$

If the two sample are independent:

$$SE(\bar{x}_2 - \bar{x}_1) = \sqrt{(SE(\bar{x}_1))^2 + (SE(\bar{x}_2))^2}$$

and $SE(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$ is estimated by $\frac{s_1}{\sqrt{n_1}}$ if sample size n1, n2 are not large, then the p-value neeed to computed from the t-distribution.

if assuming $\sigma_1 = \sigma_2$,then pooled estimate for $\sigma_1 = \sigma_2$,given by

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Paired-difference test the independent assumption is in the sampling of the couples H0: population difference is zero $t = \frac{\bar{d} - 0}{SE(\bar{d})}$, where d_i is the difference of the ith couple. $SE(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$, estimate σ_d by s_d

The sign test

4.4 第四章假设检验之卡方检验——类别变量研究

Testing of goodness-of-fit: 研究一个分类变量的分布和已知分布是否一致。H0: the color distribution is given by that table

$$\chi^2_{n-1} = \sum_{all\ categories} \frac{(observed - expected)^2}{expected}$$

期望值来自于已知分布

Testing homogeneity: $\chi^2 - test\ of\ homogeneity$ tests that the distribution of a categorical variable(color) is the same for serval populations(milk, peanut,caramel); 检验一个分类变量在不同的总体中的分布是否一致。

$$\chi^2(no.\ of\ columns - 1)(no.\ of\ rows - 1) = \sum_{all\ \ cells} \frac{(observed - expected)^2}{expected}$$

期望值来自于把不同总体合并为一个总体，计算该分类变量在所有总体中概率分布。

Testing independence computed exactly as in the case of testing homogeneity. 比较图（待粘贴）

4.5 第五章假设检验之 F 检验——方差分析

通过比值比较组内差异和组间差异的大小。Compare the sample variance of the means to the sample variance within the groups. Analysis of Variance(ANOVA)

k groups and the jth group has n_j observations:

There are total $N = n_1 + ... + n_k$ observations. Sample meen of jth group :

$\bar{y_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$

Overall sampeld mean:

$\bar{y} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}$

The treatment sum of squares :

$$SST = \sum_j \sum_i (\bar{y_j} - \bar{y})^2$$

has k-1 degrees of freedom.

The treatment mean square:

$$MST = \frac{SST}{k-1}$$

Measures the variability of the treatment mean $\bar{y_j}$

The error sum of suqares :

$$SSE = \sum_j \sum_i (\bar{y}_i j - \bar{y_j})^2$$

has $N-k$ degress of freedom. the error mean mean square :

$$MSE = \frac{SSE}{N-k}$$

Measures the varibility within the groups.

Compare the variation between the groups to the varition within the groups:

$$F = \frac{MST}{MSE}$$

follows F-distribution with k - 1 and N - K degress of freedom. under null hypothesis ,it shoud be close to 1(not exactly for chane error.)

the ANOVA table: (待粘贴)

The one-way ANOVA model:

$$y_{ij} = \mu_j + \epsilon_{ij} (\mu_j : mean\ of\ jth\ group, \epsilon_{ij} \sim N(0, \sigma^2))$$

so the null hypothesis:

$$\mu_1 = \mu_2 = ...\mu_k$$

group mean's devation away from the overall mean: $\tau_j = \mu_j - \mu$

so the the model:

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

where τ called treatment effect of group j. Then the null hypothesis is

$$H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0$$

estimate overall mean μ by the 'grand mean' $\bar{y}$, then the estimate of $\tau_j = \mu_j - \bar{y}$. the estimate of ϵ is the residual $y_{ij} - y_j$

corresonding to the model $y_{ij} = \mu + \tau + \epsilon_i j$ wecan write y_{ij} as the sum of the corresponding estimates:

$$y_{ij} = \bar{y} + (\bar{y_j} - \bar{y}) + (y_{ij} - \bar{y_j})$$

it turns out that such a decomposition is also true for the sum of squares:

$$\sum_j \sum_i (y_{ij} - \bar{y})^2 = \sum_j \sum_i (y_j - \bar{y})^2 + \sum_j \sum_i (y_i j - bary_j)^2$$

$$TSS = SST + SSE$$

it split the total variation into weo 'sources': SST and SSE.

MORE: The F-test assumes that all group have the same σ^2, it can be roughly checked with side-by-side boxplots, and there are also formal test. another assumption: data are independent within and across group. it would be the case if the data were assigned to treatment at random . F-test give conclusion not eual,how thy differ, examine all paires of means of a two sample t-test using $s_{pooled} = \sqrt{MSE}$, multiple tests, adjustment is necessary such as Bonferroni adjustment.

4.6 第六章假设检验之多重假设检验

p value $< 1\% \rightarrow$ test is 'highly significant' interpretation: If there is no effect, then there is only 1% chance to get such a highly significant result.

but if we do 800 tests, then enen their id no effect at all we expect to see 800* 1% = 8 highly results just by chance.

this is called multiple testing fallacy or look-elsewhere effect.(leads to data snooping or in other words, data dredging.)

Data snooping and other problems have lead to a crisis with regard to replicability (getting similar conclusions with different samples, procedures and data analysis methods) and reproducibility (getting the same results when using the same data and methods of analysis.)

Bonferroni correction: If there are m tests, multiply the p-values by m

False Discovery Proportion (FDP):

$$FDP = \frac{number\ of\ false\ discoveries}{total\ number\ of\ discoveries}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

False discovery rate (FDR): Controls the expected proportion of discoveries that are false. Benjamini-Hochberg procedure to control the FDR at level = 5% (say): 1. Sort the p-values: p(1) . . . p(m) 2. Find the largest k such that p(k) k m α 3. Declare discoveries for all tests i from 1 to k

Using a validation set: Split the data into a model-building set and a validation set before the analysis

5 第五篇从演绎到推断——计算机重抽样

Monte Carlo Method: 从总体中多次抽样，计算多个样本的统计量，得到样本统计量的均值与 SE 的估计。Bootstrap: plug-in principle, 从一个样本中多次放回抽样计算多个样本统计量，得到样本统计量的均值与 SE。Nonparametric bootstrap parametric bootstrap, bloc bootstrap.

bootstrap confidence interval bootstrapping for regression