Measure of central tendency

definition

1. mean: arithmetic average of a set of numbers, or distribution.
2. median: the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half.
3. mode: the value that appears most often in a dataset

when to use median

1. dataset is skewed.
2. dealing with ordinal data.

when to use mode

1. dealing with nominal data.

advantage of mean

1. easy to compute and update
2. easy to interpret
3. uses every value in the data, is a good representative of the data
4. closely related to standard deviation
5. Repeated samples drawn from the same population tend to have similar means. The mean resists the fluctuation between different samples. (central limit theory)

Measure of variability/Spread

sample variance (the $\hat{\sigma}$ estimated by sample)

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ Var\left(\widehat{\sigma^2}\right) = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

the mean and variance of sample mean: $\mu, \sigma/\sqrt{n}$

median of two data set: $median_{min}$ (#m<), $median_{max}$ (<#n), delete m+n

median variance: suppose there are 1000 data points, sample 100 data points randomly, calculate its median, do this for 1000 times, then we get the distribution of median.

Overfitting

check: have a good/bad performance in training/testing data.

reason

1. sample cannot provide overall information: recognize cats, black cats
2. too much noise, cannot see true features: abnormal cats
3. model is too complex, try to capture unnecessary features so that the model cannot be generalized: consider sleeping time

remedy

1. collect more data/data augment to have more information
2. regularization to prevent over learning
3. control the complexity of model, Occam's Razor
4. reduce #feature: feature selection, PCA, correlation matrix
5. early stop: when the error in testing data increases, stop learning

Hypothesis testing: inference from sample to population

sample size:

1. Treatment effect: $n = \frac{\sigma^2}{\Delta^2}\left(Z_{\alpha/2} - Z_{\beta}\right)^2$ [variance, difference, α, β]
2. Ratio change: target value, base value, α, β

***p*-value**: the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

type I/ II error

The type I error rate or significance level is the probability of rejecting the null hypothesis given that it is true.

type I error is the rejection of a true null hypothesis, while a type II error is the non-rejection of a false null hypothesis

power: 1-β

confidence interval

For a known standard deviation: $\left(\bar{x} - z^*\frac{\sigma}{\sqrt{n}}, \bar{x} + z^*\frac{\sigma}{\sqrt{n}}\right)$

For an unknown standard deviation:$\left(\bar{x} - t^*\frac{s}{\sqrt{n}}, \bar{x} + t^*\frac{s}{\sqrt{n}}\right)$

How to narrow: increase sample size, reduce variability,
use on-side interval, lower confidence level.

effect size: 差异类: Cohen's d, 相关类: η^2

test the distribution: Goodness of fit tests

1. Likelihood Ratio Test:$LRT = -2ln\frac{p(data|p = MLE)}{p(data|H_0)}$
2. χ^2 test

Theory

<u>中心极限定理：</u>无论总体分布如何，在满足某些条件时（e.g., 取样次数够多，样本量很大），样本均值服从正态分布。

in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.

<u>大数定律：</u>抽样次数多了，样本均值收敛/依概率收敛于真值。

the sample average converges in probability towards the expected value

the sample average converges almost surely to the expected value

Tests

<u>F test:</u>

1. Multiple-comparison ANOVA problems

The formula for the one-way ANOVA F-test statistic is

$$F = \frac{\text{explained variance}}{\text{unexplained variance}},$$

or

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}.$$

The "explained variance", or "between-group variability" is

$$\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2 / (K-1)$$

The "unexplained variance", or "within-group variability" is

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 / (N-K)$$

2. Regression problems: does the variance decrease after prediction?

χ^2 test

1. independence in contingency tables

2. goodness of fit of observed data to hypothetical distributions $\chi^2 = \sum \frac{(A-T)^2}{T}$

A: actual value, T: theoretical value

e.g. 一个骰子是不是均匀的

Z test

Mean difference: attest if sample with $\bar{x}$ is dif from known N(μ, σ): $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

T test

Mean difference: attest if sample with $\bar{x}$ is dif from $N(\mu, \sim)$

$$z = \frac{\bar{x}-\mu}{s/\sqrt{n}},\ s = \sum(x_i - \bar{x})^2/(n-1), df = n-1$$

K-S test

$D_n = \sup_x |F_n(x) - F(x)|$, the supremum of the distances set.

Common question:

1. Single sample belongs to population? $N\sim(\mu, \sigma)$

test μ:

σ known: z test

σ unknown: t test, $df = n-1$

test σ:

μ known: $\chi^2 = \frac{\sum(x_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$

μ unknown: $\chi^2 = \frac{\sum(x_i-\bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$

2. Compare two normal distribution $N_1 \sim (\overline{x_1}, s_1)$, $N_2 \sim (\overline{x_2}, s_2)$

same mean: t test:

$$t = (\overline{x_1} - \overline{x_1})/s \sim t(n_1 + n_2 - 2)$$

$$s^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-1}\right]$$

same variance: F test

$$F = \frac{s_1^2}{s_2^2} = \frac{1/m\sum_{i=1}^{m}(x_{1i} - \overline{x_1})^2}{1/n\sum_{i=1}^{n}(x_{2i} - \overline{x_2})^2} \sim F(m-1.n-1)$$

Distributions

Riht skewed distribution

Example: χ^2, F(df samll), geometric, exponential

How to compare two skewed distribution:

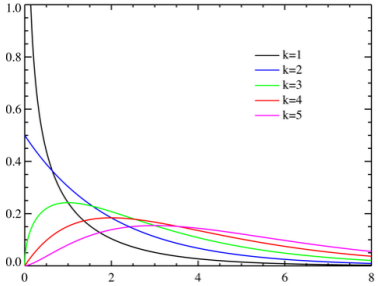
1. Distribution-Free/Non-parametric methods: Mann-Whitney test

2. log-transformation

χ^2 distribution with df = k is the distribution of a sum of the squares of k independent standard normal random variables. If $Z_1, \dots Z_k$ are independent, $\sim N(0,1)$, then

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2.$$

χ^2检查是不是属于同一分布, one-tail: if Z_i 属于同一个分布 (can be scaled to standard normal distribution using same scalars), 则χ^2>criteria value 的可能性很小，→if χ^2>criteria value, then Z_i are not from same distribution

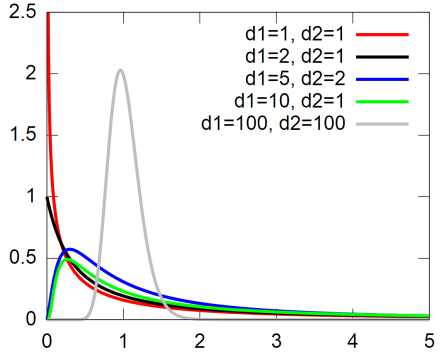


***F* distribution**: If $X \sim N(0,1)$, $(X_1, X_2 \dots, X_{n1})$ and $(Y_1, Y_2, \dots, Y_{n2})$ are two independent samples from X.

$$F = \frac{\sum_{i=1}^{n_1} X_i^2}{n_1} / \frac{\sum_{i=1}^{n_2} Y_i^2}{n_2} \sim F(n_1, n_2)$$

F 两个分布的方差是不是一样

F和χ^2的区别: F对于两种分布各自规约，方差相同即可； χ^2同时规约，要求是同一个分布



Normal Distribution:

1. $X, Y \sim N(\mu, \sigma)$, $X + Y \sim N$?
   K-S test, S-W test: if NOT significant, then it is normal
   If X, Y are independent, then $X + Y \sim N$, else not necessarily normal.
   $X \sim N(\mu, \sigma)$, Y = m* X, $m \sim Bernoulli(0.5)$, then X+Y is not normal
2. Gaussian Process VS. multivariate normal
   MVN is distribution over vectors, joint distribution of finite Gaussian distribution; GP is distribution over functions, joint distribution of infinite Gaussian distribution. They are the same within any finite interval

Binomial Distribution: n 个独立试验中成功次数的概率分布
实验 n 次, 每种情况发生的次数的概率分布

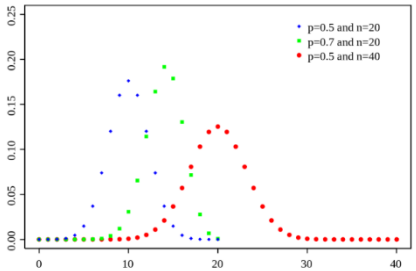
$f(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, $n \geq x \geq 0$

$E[X] = np, Var[X] = npq$.

1. 当 n 趋于∞，p 趋于 0，而 np 固定于 λ > 0，或至少 np 趋于 λ > 0 时，二项分布 B(n, p)趋于期望值为 λ 的泊松分布。

2. 当 n 趋于∞而 p 固定时，$\frac{X-np}{\sqrt{np(1-p)}}$的分布趋于期望值为 0、方差为 1 的正态分布。

3. 如果 X ~ B(n, p)和 Y ~ B(m, p)，且 X 和 Y 相互独立，那么 X + Y 也服从二项分布；它的分布为$X + Y \sim B(n + m, p)$



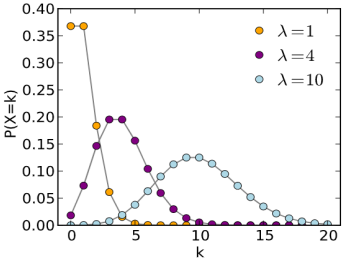
Poisson Distribution: 单位时间内随机事件发生的次数的概率分布
实验无数次(直到单位时间)，每种情况发生的次数的概率分布

$X \sim P(\lambda)$, $P(X = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$

λ是单位时间内随机事件的平均发生次数，$n \geq x \geq 0$

$E[X] = Var[X] = \lambda$.

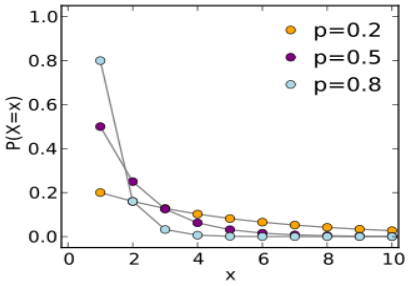
1.$X \sim Poisson(\lambda_1), Y \sim Poisson(\lambda_2)$, then $X + Y \sim Poisson(\lambda_1 + \lambda_2)$



Geometric distribution:得到一次成功所需要的試驗次数 X
$P(x) = (1-p)^x$p, p 是每次试验的成功概率

$F(p) = 1 - (1-p)^{x+1}$,

$E[X] = \frac{1}{p}$ $Var[X] = \frac{1-p}{p^2}$,

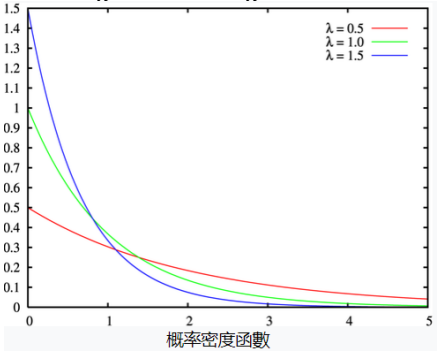


Exponential Distribution: 独立随机事件发生的时间间隔 X 的概率分布
比如旅客进入机场的时间间隔、打进客服中心电话的时间间隔。
在成功之前要实验失败次数 X，不过实验是连续不断，无穷的。

$f(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$ $F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$

λ是单位时间（或单位面积）内随机事件的平均发生次数。

$E[X] = \frac{1}{\lambda}, Var[X] = \frac{1}{\lambda^2}$.



Cross Validation

Goal: assess how the result of a statistical analysis will generalize to an independent data set.

Idea: split data into two parts, training part and testing part. Use the training data to train a prediction model, when the training is done, test the performance of the learned model on the testing data.

Type: according to how we split the data and reuse the data: holdout, leave-one-out, K-fold method.

Outlier Detection

Point/time series/context

Label data: classification and regression trees/neural network,
STL decomposition, ARIMA/exponential smoothing

Un-labell data: model: z test, box plot, Grubb test
Density: reachability distance
Distance: cluster, KNN (weak)
Robustness: isolation forest (ensemble learning)
Curse of dimensionality (better representation):
Dimension reduction: feature selection, PCA
Subspace: feature bagging

Dimension reduction

1. Feature selection: missing value ratio, low variance filter,
high correlation filter, random forest
2. Dimensionality reduction: (1) components based: FA, PCA, ICA
(2) Projection based: ISOMAP

Bayesian & Frequency

They think about parameter space in different way. Frequency stat doesn't care about all the details in parameter space, they believe the observed data is generated from a specific value in the space, so they have confidence interval. Bayesian stat cares about all the details in the parameter space, they think all the values in the parameter space can be the true value, but the probability of being true value is different, so they consider prior distribution and posterior distribution. One example to clarify the difference is that suppose our posterior distribution is binodal distribution, frequency stat will select the one with higher possibility, but Bayesian will report two peaks and corresponding probability.

Linear Regression

Assumption

1. Linearity and additive
   a) Scatter plot
   b) Non-linear transformation
   c) Fail to capture the trend
2. Predictors independent: no multicollinearity
   a) Correlation matrix, VIF (variance inflation factor), factor analysis
   b) Select important variable; stepwise regression
   PCA; regularization; add more samples
   c) X^TX singular → new predictor ≠ new information; arbitrary result
   Hard to evaluate the contribution of each predictor
   The standard error of slope parameter increases
3. Error terms distributed normally
   a) qq-plot, fit goodness test (K-S test)
   b) unusual data check
   c) CI too narrow/wide
4. Error terms constant variance
   a) Scatter plot, G-Q test
   b) Weighted Least Square, log transformation
   c) CI too narrow/wide
5. No autocorrelation, error term iid
   a) Durbin-Watson statistic
   b) Generalized difference method (OLS for change)
   c) Underestimate standard error

Q: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, x_1, x_2 are corelated, $y = \beta_3 + \beta_4(x_1 + x_2)$
$\beta_0 == \beta_3, \beta_4 = (\beta_1 + \beta_2)/2$

Q: copy all data:
coefficient: same; coefficient variance: decrease
violate observation iid

Q: 该显著的变量不显著: 依次检查假设, 样本量

How to evaluate fit goodness/model validation:

1. Residual: RMSE, MAE
2. Variance explained: R^2 (determination coefficient):
   Method 1: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = r_{y\hat{y}}^2$
   Method 2: $R^2 = \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r_{x_1x_2}^2}$
   Model comparison*: $F = \frac{(R_n^2 - R_o^2)/(k_n - k_o)}{(1-R_n^2)/(N-K_n-1)}$
3. F test $F = \frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} = \frac{R^2/k}{(1-R^2)/(N-k-1)}$
4. χ^2 test
5. Cross Validation

How to evaluate coefficient ($H_0: \beta_i = 0$)

$\because \widehat{\beta_i} = \frac{SS_{xy}}{SS_{xx}} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{12}^2}$, $\quad \therefore \mathrm{Var}(\widehat{\beta_i}) = \sigma / \sum (x_i - \bar{x})^2$

t test: $t = \frac{b}{S_b}, df = N - k - 1$

Bias and variance tradeoff

$$E\left[(f - \hat{f})^2\right] = \sigma + Var[\hat{f}] + E[f - \hat{f}]^2 = \sigma + Var[\hat{f}] + Bias[\hat{f}]^2$$

What: the property of a set of predictive models where models with higher bias in estimation have a higher variance of estimation across samples.

Bias: error from erroneous assumptions in the algorithm (underfitting)

Variance: error from sensitivity to small fluctuations in the training data. (overfitting)

Irreducible error: error from noise in the problem itself.

In regression: few parameters/simple: high bias + low variance

Parameter estimation

1. OLS: a statistical method used to find a line of best fit by minimizing the sum of squares created by a mathematical function

$$L(W) = \sum (Y_i - \widehat{Y_i})^2 = (W^T X^T - Y^T)(XW - Y)$$
$$= (W^T X^T XW - W^T X^T Y - Y^T XW + Y^T Y)$$
$$= (W^T X^T XW - 2W^T X^T Y + Y^T Y)$$
$$\frac{\partial L(w)}{\partial w} = 2X^T XW - 2X^T Y = 0$$
$$W = (X^T X)^{-1} X^T Y$$

2. MLE:

$$L(w) = logP(Y|X, w) = log \prod (y_i|x_i, w) = \sum log\,(y_i|x_i, w)$$
$$\widehat{w} = argminL(w)$$

3. Unbiased, consistency, efficiency, sufficiency

Regularization Regression

Lasso: $\sum (Y_i - w^T X_i)^2 + \lambda |w|$ feature selection

Ridge: $\sum (Y_i - w^T X_i)^2 + \lambda\, w^T w$ overfitting & multicollinearity

Difference:

1. Math: L1/2 penalty term
2. Goal: multicollinearity & overfitting → Ridge
   cannot variable selection → Lasso
3. Estimation: Lasso: Gradient descent
   Ridge: MAP $(X^T X + \lambda I)^{-1} X^T Y$

$\because$ p(y) is fixed. $\widehat{w} = argmax\, p(w|y) = argmax\, p(y|w)p(w)$

if in MLE/MAP: noise is normal, prior is normal, then OLS=MLE, Ridge=MAP

Logistic Regression:

Motivation: binary/ordinary data

Model:

Two classes: $p_1(X, \theta) = \frac{1}{1 + e^{-\theta^T X}}$

Multiclass:

multinomial:(1) one vs all (2) one vs one (3) Soft-max Regression:

$$P_j(X, \theta) = \frac{e^{\theta_j^T X}}{\sum_i^k e^{\theta_i^T X}}$$

ordinal: 1 vs 2+3+4, 1+2 vs 3+4, 1+2+3 vs 4

Assumption

1. error terms are independent
2. linearity: logit function of Y & predictors
3. Y follows Bernoulli distribution

Prediction:

$$\hat{Y} = 1 \quad \text{if} \quad p_1(X, \theta) > 0.5$$
$$\hat{Y} = argmax_j \quad Pj(X, \theta)$$

Odd: $p/(1-p)$, the probability of being A over not being A

Logit: $log\left(\frac{p}{1-p}\right)$

Parameter estimation: Gradient descent for likelihood function

$$L(\theta) = \prod_i^n \left(p_1(X_i, \theta)\right)^{y_i} \left(1 - p_1(X_i, \theta)\right)^{1 - y_i}$$
$$L(\theta) = \prod P_j(X_i, \theta), \qquad j = Y_i$$

Inference (HT)

1. LRT: $\sim \chi^2(\#coefficient)$
2. Wald test: $\frac{(\hat{p} - p_0)^2}{\hat{p}(1-\hat{p})/n} \sim \chi^2(1)$
3. Score test: $U(p_0)/I(p)$, $U(p)$ is the derivative of l with respect to p
   $I(p)$ is the Fisher information

Sampling

蓄水池采样: for n^{th} data, $p(keep) = 1/n, p(useprev) = (n-1)/n$

Probability

面条题: $f(n) = \frac{1}{2n-1}\left(1 + f(n-1)\right) + \frac{2n-2}{2n-1}\left(f(n)\right)$

$$E[len] = \sum_{i=1}^{N} \frac{1}{2i - 1}$$

扔骰子(秘书问题)