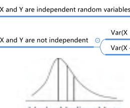
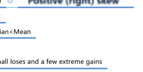
Quantitative Analysis

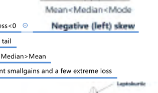
- **Probability and Distribution**
  - Independent and Dependent Event
    - P(A|B)=P(A) or P(B|A)=P(B)
  - At Least One of Two will Occur
    - P(A or B)=P(A)+P(B)-P(AB)
    - Mutually Exclusive Events: P(A or B)=P(A)+P(B)
  - Joint Probability
    - Property
      - Unconditional probability (marginal probability) is the probability of an event occurring.
      - Conditional probability, P(A | B), is the probability of an event A occurring given that event B has occurred.
    - Expression
      - P(AB)=P(A|B)×P(B)=P(B|A)×P(A)
      - For independent events: P(A|B)=P(A), So P(AB)=P(A)×P(B)
  - Bayes' Theorem
    - P(A|B)=P(B|A)P(A)/P(B)
  - Cenral moment
    - Mean and Expected Value
      - Property
        - all the observed values are summed and divided by the number of observations
        - Expected value is the weighted average of the possible outcomes of a random variable
      - Expression
        - $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$
        - E(X) = x(1)*P[X = x(1)] + ... + x(n)*P[X = x(n)]
    - Variance and Standard Deviation
      - Property
        - Provide a measure of the extent of the dispersion in the values of the random variable around the mean.
        - The square root of the variance is called the standard deviation.
      - Expression
        - defined: Var(X) = E[(X - μ)^2] = E(X^2) – [E(X)]^2, where μ = E(X)
        - If X and Y are independent random variables
          - Var(X + Y) = Var(X) + Var(Y)
          - Var(X – Y) = Var(X) + Var(Y)
        - If X and Y are not independent
          - Var(X + Y) = Var(X) + Var(Y) + 2 × Cov(X,Y)
          - Var(X – Y) = Var(X) + Var(Y) – 2 × Cov(X,Y)
    - Skewness



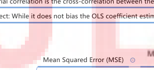
      - Positive/Right skewed
        - Skewness>0 — **Positive (right) skew**
        - Right fat tail
        - Mode<Median<Mean
        - 更受欢迎
        - Frequent small loses and a few extreme gains



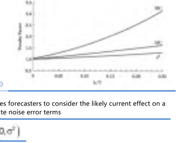
      - Nagative/Left skewed
        - Skewness<0 — **Negative (left) skew**
        - Left fat tail
        - Mode>Median>Mean
        - Frequent smallgains and a few extreme loss
    - Kurtosis



      - Leptokurtic
        - Sample Kurtosis>3
        - Excess kurtosis>0
        - 高峰肥尾
      - Mesokurtic
        - Sample Kurtosis=3
        - Excess kurtosis =0
        - Normal Distribution
      - Platykurtic
        - Sample Kurtosis<3
        - Excess kurtosis<0
  - Discrete Probability Distribution
    - Binomial Distribution
      - 均值= P（P事件的成功概率）
      - 方差=P×(1-P)
    - Poisson Distribution
      - 均值=方差= λ
  - Continuous Probability Distribution
    - Normal Distribution
      - Standard Normal Distribution: 均值=0，方差=1
      - Standardization：(observation - population mean)/standard deviation
    - Lognormal Distribution
      - If lnX is normal ,then X is lognormal
      - Right Skewed
    - Student's t-Distribution
      - As the degrees of freedom (the sample size) gets larger, the shape of the t-distribution more closely approaches a standard normal distribution.
      - Fatter tails than the normal distribution
    - Chi-Squared Distribution
      - Bounded below by zero
      - Approaches the normal distribution in shape as the degrees of freedom increase.
    - F-Distribution
      - Right-skewed and is truncated at zero on the left-hand side.
  - Central Limit Theorem
    - 条件
      - n≥30,sample statistic is a random variable
      - 总体的均值和方差已知
    - 结论
      - Sample mean服从正态分布
      - Sample mean的均值=总体均值
      - Sample mean的方差=总体方差/n
    - standard Error表示sample mean的标准差
- **Descriptive Statistics and Inferential Statistics**
  - Covariance and Correlation
    - Covariance
      - Property
        - Covariance measures the extent to which two random variables tend to be above and below their respective means for each joint realization
      - Expression
        - Cov(X,Y)=E[(X-E(X))×（Y-E(Y)]
        - Cov(X,X)=E[(X-E(X))×（X-E(X)]=σ^2(X)
    - Correlation
      - Property
        - Correlation is a standardized measure of association between two random variables
        - 只衡量线性关系，ρ=0，只表示没有线性关系
      - Expression
        - ρ=Cov(x,y)/σ(1)*σ(2)
  - Best Linear Unbiased Estimator (BLUE)
    - Unbiased
    - Efficiency
    - Consistent
    - Linear
  - Standard Error of the Sample Mean
    - 总体标准差已知 — $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
    - 总体标准差未知 — $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
  - Confidence Interval
    - the population variance is known — $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
    - the population variance is unknown — $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
  - Hypothesis Testing
    - Properties
      - a null and an alternative hypothesis.
      - The selection of the appropriate test statistic (z, t, χ2, F).
      - Specification of the significance level
      - The calculation of a sample statistic
      - A decision rule
        - With critical value
          - Reject H0 if |test statistic| > critical value
          - Fail to reject H0 if |test statistic| < critical value.
        - With p-value
          - If P-value < alpha, we reject null hypothesis
    - Two/one Tailed
      - A two-tailed test results from a two-sided alternative hypothesis (e.g., HA: μ ≠ μ0).
      - A one-tailed test results from a one-sided alternative hypothesis (e.g., HA: μ > μ0, or HA: μ < μ0).
    - Type I Error and Type II Error
      - Type I Error：错杀好人；Type II Error：放走坏人
      - P(Type I error)=α
      - Power of test（检验的势）=1-P(Type I error)
      - Type I Error and Type II Error此消彼长
      - n增大时，Type I Error and Type II Error都减小
  - Mean Hypothesis Testing
    - known population variance
      - z统计量
        - $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
        - N(0,1)
    - unknown population variance
      - t统计量
        - $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
        - t(n – 1)
  - Variance Hypothesis Testing
    - Normally distributed population
      - 卡方检验
        - $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
        - X^2(n – 1)
    - Two independent normally distributed populations
      - F检验
        - $F = s_1^2 / s_2^2$
        - F(n1 – 1, n2 – 1)
  - Copula
    - Gaussian copula
    - Student's t-copula
- **Regression Analysis**
  - One Regressor
    - 基本思想
      - 研究解释变量X与被解释变量Y的相关关系，建立这两个变量之间的数学模型
      - $Y_i = B_0 + B_1X_i + \varepsilon_i$
    - Ordinary Least Squares
      - 截距项与斜率项的估计
        - $\hat{b}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$
        - $\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}$
      - 可决系数
        - R^2 — $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$
        - r2(correlation coefficient) — $r^2 = R^2 \rightarrow r = \pm\sqrt{R^2}$
        - Adjusted R2 — $\text{Adjusted } R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$
        - SER — $SER = \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{\sum \hat{\varepsilon}_i^2}{n-2}}$
      - 三大假设条件
        - 假设均值为零
        - 独立同分布
        - 无极端异常值
  - Multiple Regressor
    - 研究解释变量X1,X2,...Xk和被解释变量Y的相关关系，建立它们之间的数学模型
    - $Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \dots + B_kX_{ki} + \varepsilon_i$
  - Joint Hypothesis Testing
    - Property
      - An F-test is used to test whether at least one slope coefficient is significantly different from zero
      - H0: b1= b2= b3= ... = bk=0; Ha: at least one bj ≠ 0 (j = 1 to k)
    - Calculation
      - Decision rule: reject H0 , if F(test-statistic) > Fc(critical value)
  - Multicollinearity
    - Multicollinearity refers to the situation that two or more independent variables are highly correlated with each other
    - Effect: OLS estimators will be computed, but the resulting coefficients may be improperly estimated (Imperfect Multicollinearity)
    - Detect: t-tests indicate that none of the individual coefficients is significantly different than 0, while the F-test indicates overall significance and the R2 is high
  - Heteroskedasticity
    - The error term u(i) is homoskedasticity if the variance of the conditional distribution of u(i) given X(i) is constant for i = 1,...,n and in particular does not depend on X(i)
    - Effect
      - The standard errors are usually unreliable estimates
      - The coefficient estimates (the b1) aren't affected



  - Autocorrelation
    - Serial correlation is the cross-correlation between the values of the same variables
    - Effect: While it does not bias the OLS coefficient estimates, the standard errors tend to be underestimated
- **Modeling and Forecasting Trend**
  - Model Selection
    - Mean Squared Error (MSE) — $MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$
    - S2 Measure — $s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k}$
    - Akaike Information Criterion (AIC) — $AIC = e^{\left(\frac{2k}{T}\right)}\frac{\sum_{t=1}^{T} e_t^2}{T}$
    - Schwarz Information Criterion (SIC) — $SIC = T^{\left(\frac{k}{T}\right)}\frac{\sum_{t=1}^{T} e_t^2}{T}$



  - SIC has the largest penalty factor
- **Modeling Cycles**
  - MA
    - The first-order moving average process enables forecasters to consider the likely current effect on a dependent variable of current and lagged white noise error terms
    - $y_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1+\theta L)\varepsilon_t; \varepsilon_t \sim WN(0,\sigma^2)$
  - AR
    - The AR(1) process seeks to explain the dependent variable in terms of a lagged observation of itself and an error term
    - $y_t = \varphi y_{t-1} + \varepsilon_t; \varepsilon_t \sim WN(0,\sigma^2)$
    - $(1-\varphi L)y_t = \varepsilon_t$
  - ARMA
    - The ARMA process incorporates the lagged error elements of the moving average process and the lagged observations of the dependent variable from the autoregressive process
    - $y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \theta\varepsilon_{t-1} + \dots + \theta_q\varepsilon_{t-q} + \varepsilon_t; \varepsilon_t \sim WN(0,\sigma^2)$
- **Estimating Volatilities and Correlations**
  - ARCH
    - the objective is to generate a greater influence on recent observations, then the a s will decline in value for older observations.
    - $\sigma_n^2 = \gamma V_L + \sum_{i=1}^{m}\alpha_i u_{n-i}^2$
  - EWMA
    - In an exponentially weighted moving average model, the weights assigned to the αi decline exponentially as we move back through time
    - Formulas
      - $\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)u_{n-1}^2$
      - $\sigma_n^2 = (1-\lambda)\sum_{i=1}^{m}\lambda^{i-1}u_{n-i}^2 + \lambda^m\sigma_{n-m}^2$
      - $Cov_n = \lambda Cov_{n-1} + (1-\lambda)X_{n-1}Y_{n-1}$
  - GARCH
    - Is the weighted sum of a long run-variance, the most recent squared-return, and the most recent variance
    - The higher the persistence (α + β), the longer it will take to revert to the mean
    - Formulas
      - $\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta\sigma_{n-1}^2$
      - $\omega = \gamma V_L$
      - $\gamma + \alpha + \beta = 1 \quad \alpha + \beta < 1$
      - $V_L = \frac{\omega}{1-\alpha-\beta}$
      - $cov_n = \omega + \alpha x_{n-1}y_{n-1} + \beta cov_{n-1}$
- **Simulations**
  - Monte Carlo Simulation
    - Specify the data generating process
    - Estimate an unknown variable
    - Save the estimate from step 2
    - Go back to step 1 and repeat this process N times
  - Variance Reduction Technique
    - Antithetic Variates
    - Control Variates
    - Bootstrapping

融跃财经
RONGYUE FINANCE