

Quantitative

Probability concepts

1. A random variable is a quantity whose outcome is uncertain.
2. Probability is a number between 0 and 1 that describes the chance that a stated event will occur.
3. An event is a specified set of outcomes of a random variable.
 1. Mutually exclusive events(互斥事件) can occur only one at a time.
 2. Exhaustive events(遍历事件、穷举事件) cover or contain all possible outcomes.
 3. Independent events(独立事件)
 4. Dependent events(相依事件)
4. 概率的两个属性 (The two defining properties of a probability)
 1. $0 \leq P(E) \leq 1$ (Where $P(E)$ denotes that probability of an event E)
 2. that sum of the probabilities of any set of mutually exclusive and exhaustive events equals 1.
5. Types of probability
 1. Empirical probability 经验概率
 - A probability estimated from data as a relative frequency of occurrence is an empirical probability.
 - Estimate the probability of an event based on **historical data**.
 2. Priori probability 先验概率
 - A probability obtained based on **logical analysis**逻辑分析 is an priori probability.
 3. Subjective probability
 - Drawing on **personal or subjective judgement** without reference to any particular data.
6. Odds
 1. Odds for event E: $E = P(E) / [1 - P(E)]$
 2. Odds against event E: $E = [1 - P(E)] / P(E)$
7. Unconditional & conditional probability
 1. Unconditional probability: A probability of an event not conditioned on another event, denoted $P(A)$.
 2. Conditional probability: Probability of an event(A) conditioned another event(B), denoted $P(A|B)$.
 $P(A|B) = P(AB)/P(B)$, $P(B) \neq 0$. 在B发生的条件下, A发生的概率。
8. Joint probability 联合概率 and multiplication rule 乘法规则
 1. The probability of both A and B occurring is the joint probability of A and B, denoted $P(AB)$.
 2. The multiplication rule of probabilities is $P(AB) = P(A|B) * P(B)$.
9. Probability that at least one of two events occur, Denoted as $P(A \text{ or } B)$, or $P(A+B) = P(A) + P(B) - P(AB)$
 - If event A and B are mutually exclusive, then $P(A \text{ or } B) = P(A + B) = P(A) + P(B)$
10. Total probability rule
 1. According to the total probability rule, if S_1, S_2, S_3, \dots are mutually exclusive and exhaustive scenarios or events, then $P(A) = P(A|S_1)P(S_1) + P(A|S_2)P(S_2) + \dots + P(A|S_n)P(S_n)$
11. Bayes' formula
 - $P(A|B) = \frac{P(B|A)}{P(B)} * P(A)$

- PS: Total probability rule always works with Bayes' formula.
- Bayes' formula is a method for updating probabilities based on new information.
- Bayes' formula is expressed as follows:
 - Updated probability of event given the new information = [(Probability of the new information given event)/(Unconditional probability of the new information)] * Prior probability of event

12. Probability weighted expected Value E(X)

1. The expected value of a random variable is a probability-weighted average of the possible outcomes of random variable. For a random variable X, the expected value of X is denoted E(X).
- $E(X) = P(x_1)x_1 + P(x_2)x_2 + \dots + P(x_n)x_n$

Total probability rules for expected value:

- $E(X) = E(X|S_1)P(S_1) + E(X|S_2)P(S_2) + \dots + E(X|S_n)P(S_n)$

13. The variance(σ^2) of a random variable

1. The variance of a random variable is the expected value (the probability-weighted average) of squared deviations from the random variable's expected value E(X):

$$\text{公式: } \sigma^2(X) = P(x_1)[x_1 - E(X)]^2 + P(x_2)[x_2 - E(X)]^2 + \dots + P(x_n)[x_n - E(X)]^2$$

2. Variance is a measure of dispersion about the mean. Increasing variance indicates increasing dispersion. Variance is measured in squared units of the original variable.
3. Standard deviation is the positive square root of variance. Standard deviation measures dispersion, but it is measured in squared units of the original variable.

14. 协方差Covariance

1. Covariance is a measure of the co-movement between random variables.
 1. **The covariance of a random variable with itself is its own variance.**
2. **Positive covariance** 同涨同跌
3. **Negative covariance** 你涨我跌

$$\text{4. 公式: } Cov(R_A, R_B) = \sum_{i,j=1}^n P(R_{A,i}, R_{B,j})[(R_{A,i} - E(R_A))][R_{B,j} - E(R_B)]$$

5. Disadvantage of covariance:

1. Values range from minus infinity (负无穷) to positive infinity (正无穷).
2. Units of covariance difficult to interpret.

15. 相关系数Correlation

1. 定义: A standardized measure of **linear relationship** (linear association 线性关联) between two variables.

$$\text{2. 公式: } \rho_{i,j} = \frac{Cov(R_i, R_j)}{\sigma_i \sigma_j}$$

3. 相关系数的含义:

- If two variables have a very strong (inverse) linear relation, then absolute value of their correlation will be close to 1 (-1).
- If two variables have a weak linear relation, then the absolute value of their correlation will be close to 0.

- Values range from + 1(perfect positive correlation) to -1 (perfect negative correlation)

16. Labeling

17. Combination 组合

- The number of ways that we can choose r objects from a total of n objects,when the order in which the r objects are **listed does not matter**.
- C_{10}^5
 - Calculator: [10] [2nd] [+] [5] 计算出来的结果是: 252

18. Permutation 排列

- The number of ways that we can choose r objects from a total of n objects,when the order in which the r objects are listed **does matter**.
- P_{10}^5
 - Calculator: [10] [2nd] [-] [5] 计算出来的结果是: 30240

Common Probability Distributions

1. Probability Distributions 概率分布

1. Discrete distribution 离散分布

1. Discrete random variable(离散随机变量) takes on at most a finite and countable number of possible values.

2. Continuous distribution 连续分布

1. Continuous random variable(连续随机变量) takes on an infinite(无限的) and uncountable (不可数的) number of possible values.

2. Probability function 概率函数

1. Probability function 概率函数

- 定义: The probability function specifies the probability that the random variable will take on a specific value.
- Example: $P(X=x)$ is the probability that a random variable X takes on the value x.
- 区别:
 1. The probability function is denoted $p(x)$ for a **discrete random variable**.
 2. The probability function is denoted $f(x)$ for a **continuous random variable**.

2. Probability density function, $f(x)$ 概率密度函数

- 定义: Specifies the probability that the continuous random variable take on a value **within a range**. 概率密度函数研究的是一个区间。
- The probability of taking on a specific value is always zero, $P(X=x) = 0$.

3. Cumulative probability function, $F(x)$ 累计概率函数

- 定义: The cumulative distribution function, denoted $F(x)$ for **both continuous and discrete random variables**, gives the probability that the random variable is less than or equal to x.
- 公式: $F(x) = P(X \leq x)$
- $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

3. Discrete uniform distribution

1. 定义: Has a finite(有限的) number of possible outcomes, all of which are equally likely (同样可能) .

2. Example:

1. $P(X) = 0.2$, for $X = \{1, 2, 3, 4, 5\}$; then:

- $P(2) = 20\%$
- $F(3) = P(X \leq 3) = 60\%$ 由于 $F(x)$ 为累计概率函数
- $P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = 60\%$

4. Bernoulli random variable (Bernoulli trial)

1. 定义:

1. The probability of success (p) is constant for all trials. $P(X=1) = p$
2. The probability of failure ($1-p$) is constant for all trials. $P(X=0) = 1-p$
3. The trials are all independent.

2. 期望: Expected value for binomial random variable = np

3. 方差: Variance for binomial random variable = $np(1-p)$

4. The probability of binomial random variable:

- 公式: $P(x) = C_n^x p^x (1-p)^{n-x}$
- 抛硬币或者投篮, 比如抛10次, 有6次朝上的概率计算。

5. Binomial tree

A binomial tree is the graphical representation of a model of asset price dynamics in which, at each period, the asset moves up with probability p or down with probability $(1-p)$. The binomial tree is a flexible method for modeling asset price movement and is widely used in pricing options.

6. Continuous uniform distribution 连续的均匀分布

1. Probability density function 概率密度函数

- $f(x) = \frac{1}{b-a}$ For $a < x < b$
- $f(x) = 0$

2. Cumulative probability distribution

- $F(x) = 0$ For $x \leq a$
- $F(x) = \frac{x-a}{b-a}$ For $a < x < b$
- $F(x) = 1$ For $x \geq b$

7. Normal distribution

1. 定义: The normal distribution is a continuous symmetric probability distribution that is completely described by two parameters: **its mean, μ , and its variance, σ^2 .**

2. 属性:

- Completely described by mean and variance.
- Skewness = 0
- Kurtosis = 3
- Linear combination of normally distributed random is also normally distributed.
- Probabilities decrease further from the mean, but the tails go on forever.

3. Probability for given interval

- 68% confidence interval = $\mu \pm 1\sigma$
- 90% confidence interval = $\mu \pm 1.65\sigma$
- 95% confidence interval = $\mu \pm 1.96\sigma$
- 99% confidence interval = $\mu \pm 2.58\sigma$

8. Standard normal distribution

1. 定义: The standard normal random variable, denoted Z , has a mean equal 0 and variance equal to 1.
2. All questions about any normal random variable can be answered by referring to the cumulative distribution function for a standard normal random variable, denoted $N(x)$ or $N(z)$.
3. 标准化: $Z = \frac{X - \mu}{\sigma}$
4. Example:
 - 1.

Related Questions: Quantitative Methods (2)

Q. A normally distributed random variable has a mean of 100 and a standard deviation of 12. The probability of observing a value greater than 82 is the cumulative distribution function (cdf) of the standard normal variable:

- A. $N(1.5)$.
- B. $N(-1.5)$.
- C. $1 - N(1.5)$.

Solution

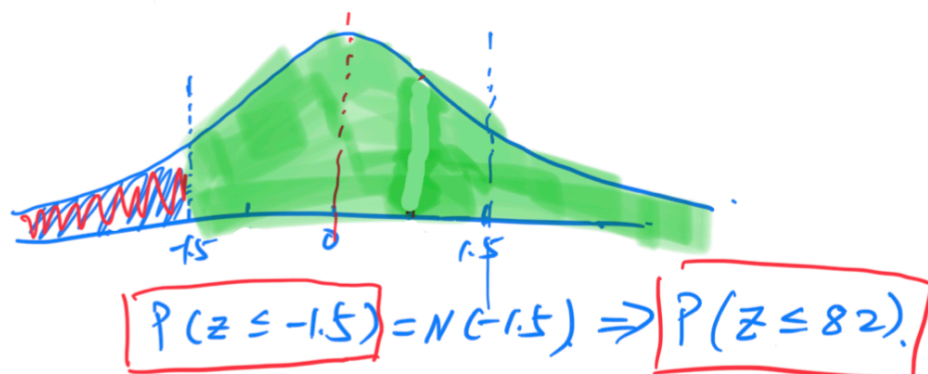
A is correct. The standardized value of this normal distribution can be obtained using the formula $(X - \mu)/\sigma = (82 - 100)/12 = -1.5$. The cdf of $N(-1.5)$ provides the probability of a value less than or equal to 82.

B is incorrect because $1 - N(1.5) = N(-1.5)$ is the probability of a value in this distribution of less than 82.

C is incorrect because the probability of observing random draw less than 82 will be $P(Z < -1.5) = N(-1.5) = 1 - N(1.5)$.

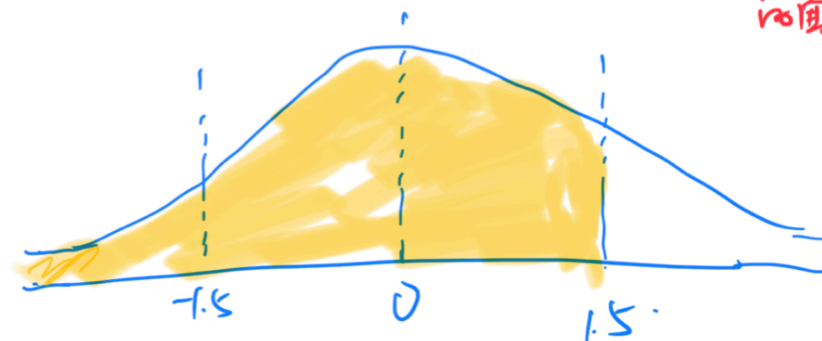
Common Probability Distributions Learning Outcome

- i. Explain how to standardize a random variable



本题求的是 $P(z > 82)$ 的概率，即绿色阴影的面积，即 $N(1.5)$ 的概率，由对称性可知。

-1.5右侧的面积。



$N(1.5)$ 的面积 = $N(-1.5)$ 右侧的面积。

9. Normal distribution: application

1. Shortfall risk

1. 定义: the risk that portfolio value or return will fall below the minimum acceptable level (R_L , shortfall level, threshold level) over some time horizon.

2. **The lower, the better.**

2. Roy's safety-first ratio

1. 定义: the distance from the mean return to the shortfall level in units of standard deviation.

2. 公式: $SFRatio = \frac{E(R_p - R_L)}{\sigma_p}$

3. The higher the better.

3. Shortfall risk vs Roy's safety-first ratio 的不同

Minimizing shortfall risk = Maximizing safety-first ratio

10. Lognormal distribution

1. x is normal distributed, then e^x is lognormal distributed.

2. Lognormal random variable is bounded from below by zero.

3. It is positively skewed.

4. It is used to model asset prices.

11. Multivariate distributions 多元分布

1. 定义: Multivariate distributions specify the probabilities for a group of related random variables. A portfolio of technology stocks represents a group of a related assets.
 2. 知识点: A multivariate normal distribution for the returns on n stocks will have n means, n variances and $n(n-1)/2$ distinct correlations.
 3. Bivariate normal distribution 二元正态分布
12. Student's t -distribution and chi-square and F -distribution
1. t -distribution
 1. 定义: The standard t -distribution is a symmetrical probability distribution defined by degrees of freedom (df) and characterized by fat tails.
 2. 参数 degrees of freedom (df): $df = n - 1$, where n is the sample size.
 3. Symmetrical (bell shaped 钟形的), skewness = 0
 4. Fatter tails than a normal distribution
 5. As df increases, t -distribution is **approaching to** standard normal distribution.
 2. Chi-square
 1. 定义: The chi-square distribution is asymmetrical (不均匀的, 不对称的), defined by degrees of freedom, and with k df is the distribution of the sum of squares of k independent standard normally distributed random variables, so it does not take on negative values.
 2. A different distribution exists for each value of df, $n-1$
 3. F -distribution
 1. 定义: The distribution is a family of asymmetrical distributions bounded from below by 0.
 2. Each F -distribution is defined by two values of degrees of freedom, the numerator df and the denominator. If x_1^2 is one chi-square random variable with m df and x_2^2 is another chi-square random variable with n df, then $F = (x_1^2/m)/(x_2^2/n)$ follows an F -distribution with m numerator df and n denominator df.
13. Monte Carlo simulation
1. 定义: Monte Carlo simulation involves the use of a computer to represent the operation of a complex financial system.
 2. Feature: the generation of a large number of random samples from specified probability distributions to produce a distribution of possible outcomes.
 3. Application
 - Used in planning
 - in financial risk management
 - Valuing complex securities
 4. Limitation
 - Fairly complex
 - Do not directly provide precise insights.
 - Provide answers no better than the assumption used.
14. Historical simulation
1. 定义: used randomly selected past data in risk factors to produce a distribution of possible outcomes.
 2. Limitation:
 - Grounded in actual data and reflect only the risks represented in the sample historical data. 仅代表历史数据支持某些结论

- Does not lend itself to "what if" analyses.

Sampling & Estimation

1. Sampling

1. Simple random sampling

1. In simple random sampling, each observation has an equal chance of being selected.

2. Stratified random sampling

1. In stratified random sampling, the population is divided into subpopulations, called strata (层级) or cells, based on one or more classification criteria (标准);
2. Simple random sampling are then drawn from each stratum.
3. Stratified random sampling ensures that population subdivisions of interest are represented in the sample.
4. Stratified random sampling produces **more-precise (更精确)** parameter estimates than simple random sampling.

3. Sampling error

1. 定义: difference between the sample statistic and the population parameter.

4. Sampling distribution

1. 定义: The distribution of all the distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population.

5. Selection of sample size

1. **Pros** for large sample size

1. a better estimate (更精确的估计)

2. **Cons** for large sample size

1. additional expense 额外的成本
2. **would not improve** the estimate for parameter

2. Sample biases 抽样偏差

1. Data-mining bias (Data snooping bias) 数据挖掘偏差: 数据挖掘无意义, 萧敬腾和“雨神”

2. Sample selection bias 样本选择偏差

Survivorship bias 幸存者偏差, 比如私募的收益率

3. Look-ahead bias 前视偏差: using information that was not available on the test date.

4. Time-period bias 后视偏差: based on a time period that may make the results time-period specific.

5. Self-selection bias 自我选择偏差: reflects the ability of entities to decide whether or not they wish to report their attributes or results and be included in databases or samples.

3. Central limit theorem 中心极限定律

1. 定义: Given a population described by **any probability distribution** having **mean μ and finite variance σ^2** , the sampling distribution of the sample mean \bar{X} , computed from samples of size n from this population, will be approximately normal with mean μ (the population mean) and variance σ^2/n (The population variance divided by n) when the **sample size n is large**.

2. 条件:

1. 简单随机抽样
2. $n \geq 30$ 大样本
3. 结论: $\bar{X} \sim (\mu, \sigma^2/n)$

4. Standard error 标准误

1. 定义: The standard deviation of the distribution of sample statistic 样本统计的标准差
2. When the population standard deviation(σ) is known:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. When the population standard deviation(σ) is **unknown**总体方差未知:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

5. Point estimate 点估计

1. 定义: The calculated value of the sample statistic in a given sample, used as an estimate of the population parameter.
2. **Estimator: formulas** to compute the **sample statistics**.
 1. 例子: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
 2. An estimator has a sampling distribution.
3. **Estimate: particular values** calculated from sample observations using an estimator.
 1. An estimate is a fixed number.
4. estimator的属性

- Unbiasedness无偏性: the expected value equals the parameter it is intended to estimate.
 $E(\bar{X}) = \mu$
- Efficiency有效性: the unbiased(公正的) estimator of the population parameter that has a sampling distribution with **smallest variance**. 在无偏的估计量中找到方差最小的。
- Consistency一致性: the probability of **estimates close to the value of the population parameter increases as sample size increases**. 随着n增大, 估计值接近总体的参数。

6. Confidence interval estimate 区间估计

1. Confidence interval for observation

1. 定义: A range for which a given percentage($1-\alpha$, called the **degree of confidence**置信度, 置信水平) of all observations will lie based on a particular probability distribution.
2. **Significance level 显著性水平(α)**: the probability that the observations would not fall in a specific range.

3. Confidence interval = Point estimate \pm Reliability factor * Standard deviation

1. **Reliability factor 可信赖因子** = a number based on the **assumed distribution** of the population and the **degree of confidence($1-\alpha$)** for the confidence interval.
2. Reliability factors for normal distribution:

1. 68% confidence intervals: $Z_{0.16} = 1$
2. 90% confidence intervals: $Z_{0.05} = 1$
3. 95% confidence intervals: $Z_{0.025} = 1$
4. 99% confidence intervals: $Z_{0.005} = 1$

4. Confidence interval for population

1. 定义: A range that contains the population mean with a given confidence(1- α) interval.
A 100(1- α) confidence interval for a parameter has the following structure:

Confidence interval of population mean = Point estimate of population mean \pm Reliability factor * Standard error of sample mean

2. 方差已知: A 100(1- α) confidence interval for population mean μ when sampling from a normal distribution with **known variance(normal distribution)** σ^2 is given by:
 - $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
 - where $z_{\alpha/2}$ is the point of the **standard normal distribution** such that $\alpha/2$ remains in the right tail.
 - We use $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ as an **alternative** to the t-distribution confidence interval for the population mean.
 - 使用z-statistic is less conservative 保守的 (narrower 狭窄的) than the corresponding confidence interval based on a t-distribution.
3. 方差未知: A 100(1- α) confidence interval for population mean μ when sampling from a normal distribution with unknown variance σ^2 (a t-distribution confidence interval) is given by:
 - $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$
 - where $t_{\alpha/2}$ is the point of the **t-distribution** such that $\alpha/2$ remains in the right tail.
 - s is the sample standard deviation
 - Because of the **central limit theorem 中心及限定理**, when dealing with a larger sample from a population with unknown variance that may be not normal.
5. "Bootstrap and jackknife are simple but powerful methods for statistical inference, and they are particularly useful when no analytical formula is available. Bootstrap constructs the sampling distribution of an estimator by repeatedly drawing samples from the original sample to find standard error and confidence interval. Jackknife draws repeated samples while leaving out one observation at a time from the set, without replacing it."

Hypothesis Testing 假设检验

1. A hypothesis is a statement about one or more populations.
 1. Addresses (解决) the questions such as "is the value of the parameter equal to a specific value"
2. Steps of hypothesis testing
 1. State the hypotheses 陈述假设.
 2. Identify the appropriate test statistic and its probability distribution 确定适当的检验统计量及其概率分布.
 3. Specify the significance level 确定显著性水平 α
 4. State the decision rule 陈述决策规则
 5. Collect the data and calculate the test statistic 收集数据并计算检验统计量.
 6. Make a decision 进行决策
3. Null hypothesis 原假设 vs. Alternative hypothesis 备择假设

1. Null hypothesis(H_0)

1. Hypothesis to be tested.
2. The "=" sign will be only in null hypothesis.

2. Alternative hypothesis(H_a)

1. The opposite side of null hypothesis.
2. Hypothesis that analyst wants to approve证明 or conclude推断.
3. Accepted when the null hypothesis is rejected.

4. Two-tailed test & One-tailed test.

Let θ indicate the true value(真实值) of the population parameter, Where θ_0 is a hypothesized value (假设值) of the population parameter.

1. Two-tailed test

1. $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$

2. One-tailed test

1. $H_0 : \theta \leq \theta_0$ versus $H_a : \theta > \theta_0$ (right side)
2. $H_0 : \theta \geq \theta_0$ versus $H_a : \theta < \theta_0$ (left side)

5. Hypothesis Testing

1. Test statistic 检验统计量

1. *Test statistic* =
$$\frac{\text{Sample statistic} - \text{Value of the population parameter under } H_0}{\text{Standard error of the sample statistic}}$$

2. 方差已知: $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

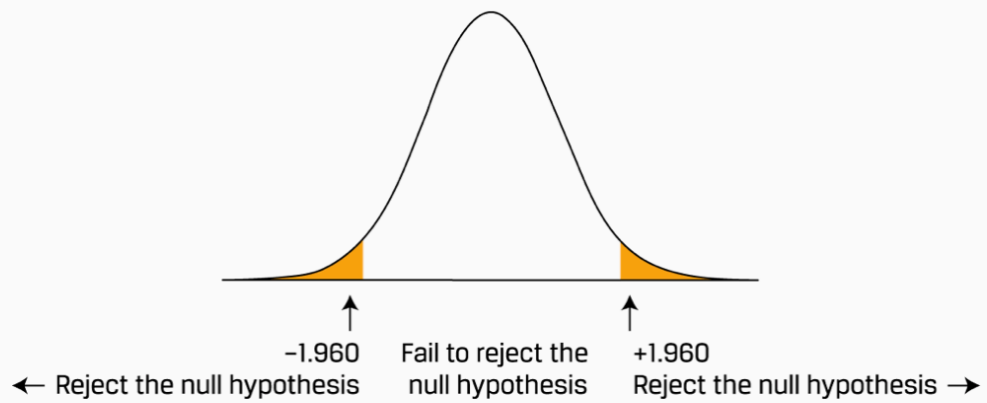
- z =z-statistic
- \bar{X} = sample mean
- μ_0 = hypothesized value of the population mean
- σ = population standard deviation

3. 方差未知: $t_{n-1} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

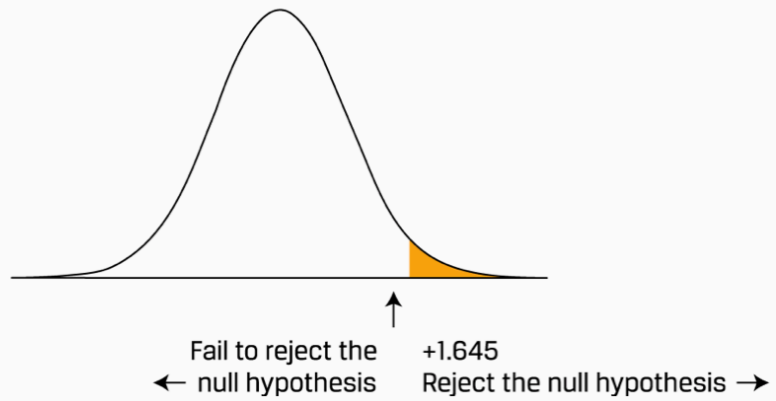
- $t_n - 1$ = t-statistic with $n-1$ df(n is the sample size)
- \bar{X} = sample mean
- μ_0 = hypothesized value of the population mean
- s = sample standard deviation

2. Critical value关键值(Rejection point拒绝点)

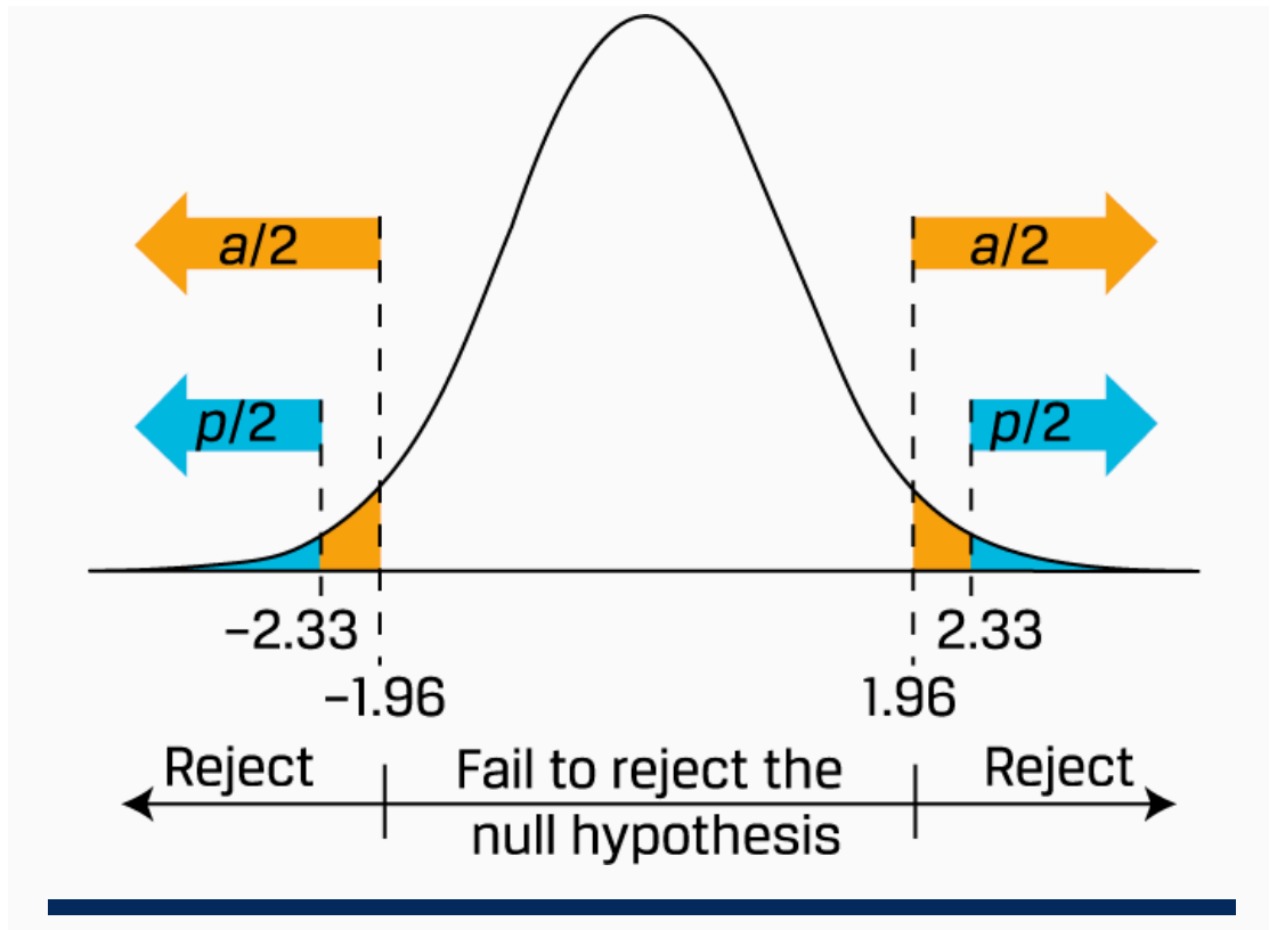
A. $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$



B. $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$



3. Significance level(α)



4. P-value p值

1. The p-value is **the smallest level of significance** at which the null hypothesis can be rejected.
2. The **smaller the p-value**, the **stronger evidence against the null hypothesis** and in favor of (有利于、赞成) the alternative hypothesis.

6. Specify the level of significance 列举显著性水平

1. **Type I Error**: rejecting null hypothesis when it is true. 拒真，错杀好人

- $P(\text{Type I error}) = \text{Significance level}(\alpha)$
- 百度解释：在进行**假设检验**时，由于**检验统计量**是**随机变量**，有一定的**波动性**，即使原假设 H_0 为真，在正常的情况下，计算的**统计量**仍有一定的**概率** α (α 称为显著性水平)落入拒绝域内，因此也有可能会错误地拒绝**原假设** H_0 ，这种当原假设 H_0 为真而拒绝原假设的错误，称为假设检验的**第一类错误**，又称为拒真错误。
- Type I error is a **false positive** 误报(reject when the null is true)
- 例子：A **level of significance of 5%** for a test means that there is a 5% probability of rejecting a true null hypothesis and corresponds(相当，相类似) to 95% confidence level.

2. **Type II error**: failing to reject the null hypothesis when it is false. 取伪，留下坏人

- $P(\text{Type II error}) = \beta$ ，通常用 β 来表示。
- Type II error is a false negative(漏报率、假阴性)(fail to reject when the null is false).

Exhibit 5. Correct and Incorrect Decisions in Hypothesis Testing

Decision	True Situation	
	H_0 True	H_0 False
Fail to reject H_0	Correct decision: Do not reject a true null hypothesis.	Type II error: Fail to reject a false null hypothesis. False negative
Reject H_0	Type I error: Reject a true null hypothesis. False positive	Correct decision: Reject a false null hypothesis.

3. Power of test: rejecting the null hypothesis when it is false. 正确的拒绝了一个错误的原假设

- Power of test = $1 - P(\text{Type II error}) = 1 - \beta$

4. 辨析:

- 第一类错误和第二类错误是此消彼长
- 随着sample size(n)增大, 第一类错误和第二类错误都会下降。

7. 假设检验的考点

- 陈述假设检验 H_0 和 H_a
- 判断何种分布
- 检验统计量
- 关键值

- z分布, 单尾、双尾?
- 已知量 $t > z$

5. 决策原则

- 检验统计量 vs 关键值 (点)
- P值 vs α (面)

8. 卡方检验- 适用于检验一个总体的方差, 而且不是双尾检验。

- In tests concerning **the variance of a single normally distributed population**, the test statistic is chi-square with $(n-1)$ degrees of freedom.

9. F-test

- For tests concerning differences between the variance of two normally distributed populations based on two random, independent samples, the appropriate test statistic is based on an F-test (the ratio of the sample variances).
- The degrees of freedom for this F-test are $n_1 - 1$ and $n_2 - 1$, where n_1 **Corresponds (相当、相类似) to the number of observations in the calculation of the numerator** 分子 and n_2 **is the**

number of observations in the calculation of the denominator 分母 of the F-statistic.

10. Parametric test and nonparametric test

1. Parametric tests 参数统计测试

1. 定义: Based on assumptions about population distribution and population parameters.
2. E.g., t-test, z-test, F-test

2. Nonparametric tests 非参数统计测试

1. Test things other than parameter values.
2. Applied when:
 - Data do not meet distributional assumptions.
 - Data are given in ranks.
 - The hypothesis being addressed does not concern a parameter.

11. Statistical significance vs. Economic significance

1. Statistical significance does not necessarily imply economic significance, due to:
 - Transactions costs
 - Taxes
 - Risk

12. Hypothesis test concerning population correlation(检查是否符合线性关系)

1. 定义: Test whether the correlation coefficient between two variables is equal to zero.
2. 假设: $H_0 : \rho = 0, H_a : \rho \neq 0$
3. T-test: $t = \frac{r\sqrt{n-2}}{1-r^2}$
 1. df = n-2;
 2. r = sample correlation coefficient
4. Two-tailed test;
5. Decision rule: reject H_0 if $t > +t_{critical}$ or $t < -t_{critical}$

Organizing, Visualizing, and describing Data (Statistical concepts and market returns)

1. Data (a statistical perspective 统计学视角)

1. numerical data
 1. continuous data
 2. discrete data
2. categorical data
 1. nominal data
 2. ordinal data

2. Data (based on how they collected) 学会辨析

1. Cross-sectional
 1. 定义: a sequence of observations for a single observational unit on a specific variable collected over time and discrete and typically equally spaced intervals of time.
2. Time series
 1. 定义: **a list of the observations of a specific variable** from **multiple observational units**

at a given point in time.

3. Pannel

1. 定义: a mixed of time-series and cross-sectional data that consists of observations through time on one or more variables for multiple observations units.

3. Data(based whether or not data in a highly organized form)

1. structured data: highly organized in a pre-defined manner, usually with repeating patterns.

1. Database

2. unstructured data: not follow any conventionally organized forms.

1. 音频、视频

4. Frequency distribution

1. Absolute frequency distribution
2. Relative frequency distribution
3. Cumulative frequency distribution

5. Quantitative descriptions of return distribution 收益分布的量化描述

1. Central tendency 集中趋势
2. Dispersion 离散
3. Skewness 偏度
4. Kurtosis 峰度

6. Measures of central tendency 集中趋势的度量

1. Mean

■ Arithmetic mean

- 定义: Arithmetic mean return focus on **average single-period performance**.
- Advantage:
 - Easy to work with mathematically.
 - Uses all the information about the size and magnitude of the observations.
- Disadvantage:
 - sensitive to extreme values

■ Geometric mean

- 定义: used to calculate average periodic **compound rate of return on investment**.
- 公式: $\bar{X}_{geometric} = \sqrt[n]{X_1 X_2 \dots X_n}$

■ Harmonic mean

- 定义: 计算平均成本
- 公式: $\bar{X}_{Harmonic} = \frac{N}{\sum_{i=1}^n \frac{1}{X_i}}$

■ 总结:

1. Harmonic Mean \leq Geometric Mean \leq Arithmetic mean
2. Harmonic Mean 比较适用于股票建仓

■ Weighted mean

- 定义: A mean in which different observations have different proportional influence on the mean.
- 公式: $\bar{X}_w = w_1 R_1 + w_2 R_2 + \dots + w_n R_n$
- 应用: 计算组合收益

2. Median: The median is the value of the middle item.
3. Mode: The mode is most frequently observed value and is the only measure of central tendency that can be used with nominal data.

7. Measure of dispersion

1. Absolute dispersion

1. Quantiles 分位数

1. Formula for location of data in ascending order(升序, 从小到大).

2. 位置计算公式: $L_y = (n + 1) \frac{y}{100}$

- n = the number of the data
- y = the y^{th} percentile

3. 位置:

1. median 中位数
2. quartiles 四分位数
3. quintiles 五分位数
4. deciles 十分位数
5. Percentiles 百分位数

4. 线性插值法

2. Range = Maximum Value - Minimum Value

3. Mean absolute deviation(MAD)

$$MAD = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

4. Variance and standard deviation

$$\text{总体方差: } \sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

$$\text{样本方差: } s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{n-1}$$

2. Relative dispersion

1. 公式: $CV = \frac{S}{\bar{X}}$

2. A measure of **risk unit of mean return**(单位回报的风险), thus the lower is the better.

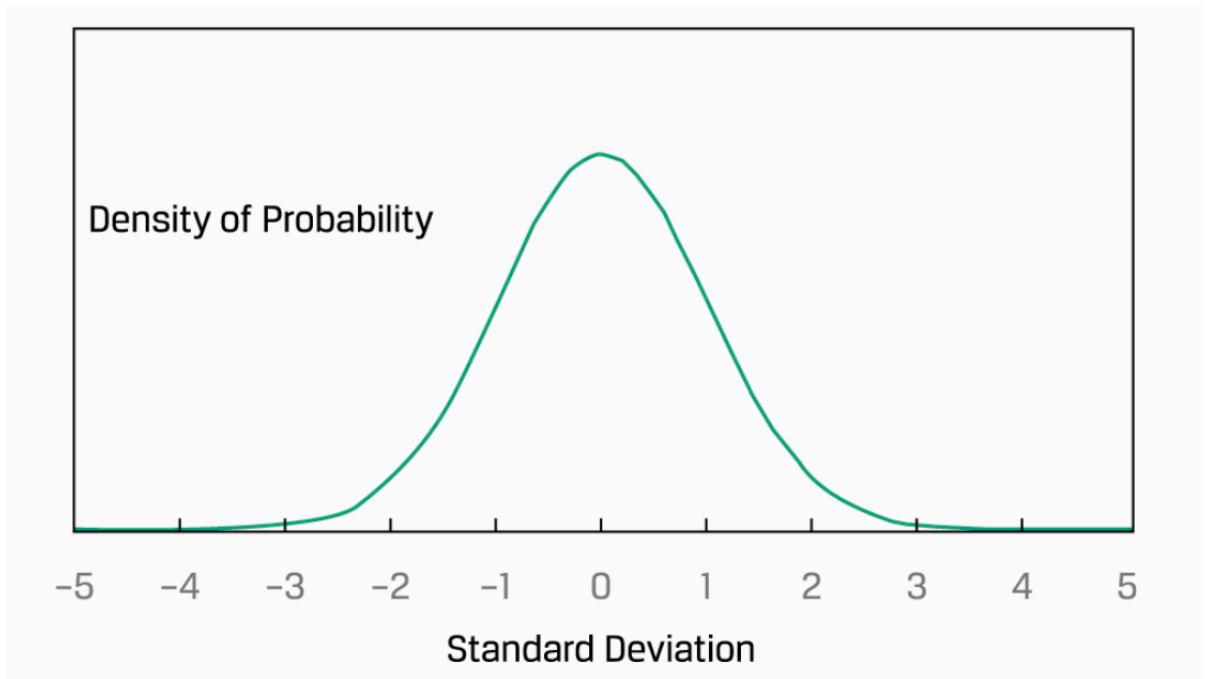
3. Chebyshev's inequality 切比雪夫不等式

1. 定义: For **any distribution** with **finite variance**, the **minimum percentage of observations** that lie within k standard deviations of the mean would be $1 - \frac{1}{k^2}$, given $k > 1$.

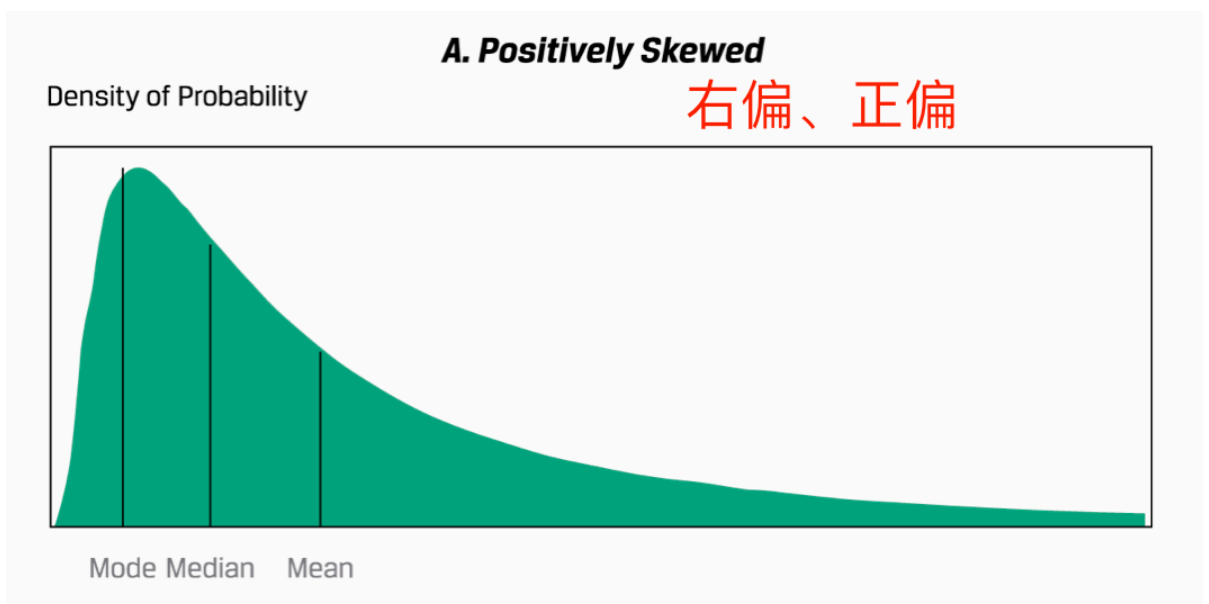
8. Skewness 偏度

1. Indicating the degree of symmetry of return distributions.

1. $S_k = 0 \rightarrow$ symmetrical distribution (bell shaped 钟形的) Symmetrical distribution 对称分布

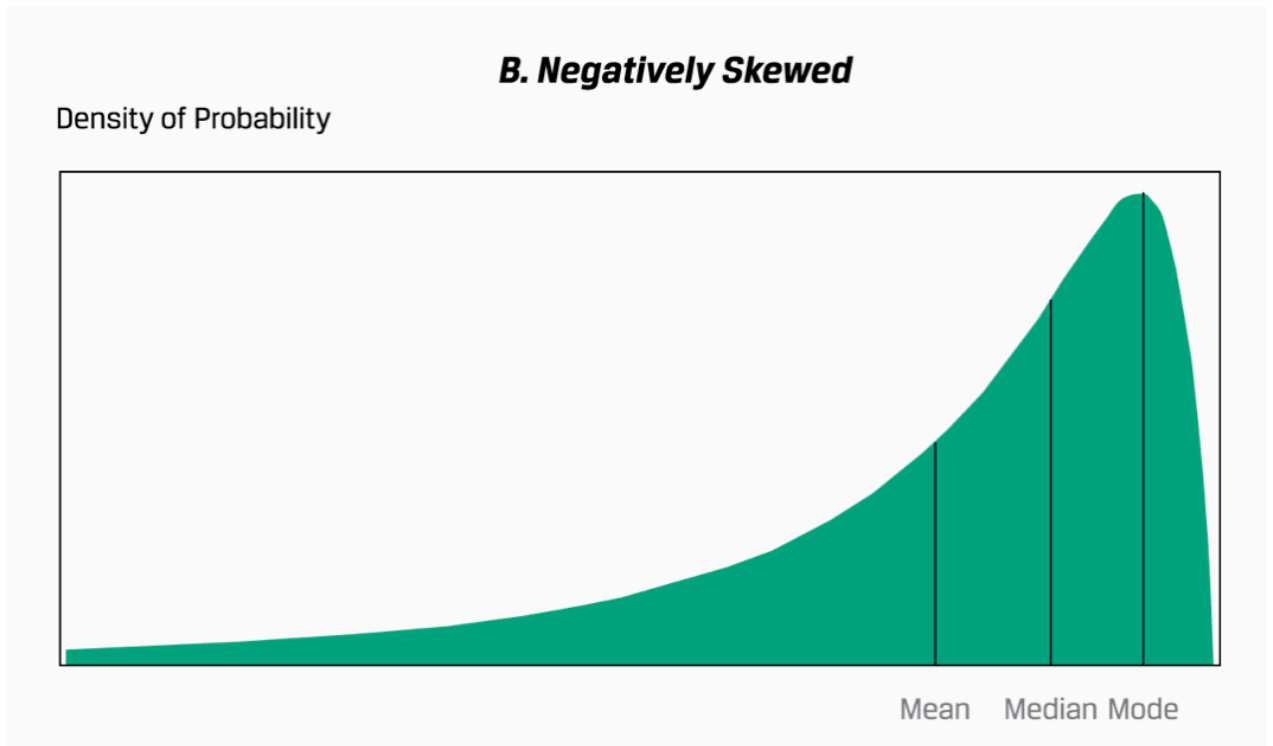


1. Mean=Median=Mode
2. It is completely described by two parameters-its **mean and variance**
2. $S_k > 0$ ->Positively (right) skewed distribution 正（右）偏态分布



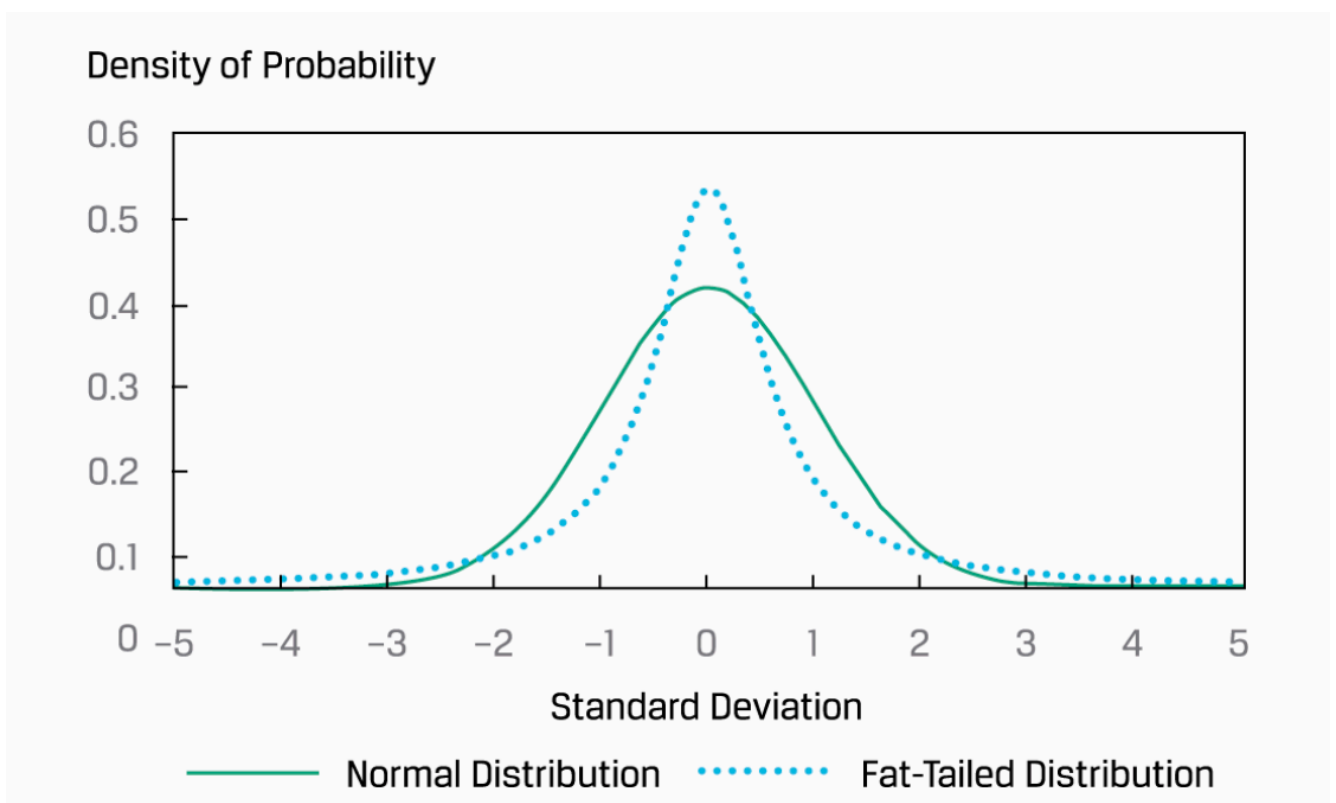
注: **Frequent small losses and a few extreme gains** 频繁的小损失和一些极端的收益

3. $S_k < 0$ ->Negatively (left) skewed distribution 正（左）偏态分布



注: Frequent small gains and a few extreme losses(fatter/longer left tail)

9. Kurtosis 峰态



1. Leptokurtic 尖峰态

1. More peaked, **fatter tailed than normal distribution.**
2. kurtosis > 3 , excess kurtosis > 0

2. Mesokurtic 常峰态

1. Identical to normal distribution
2. Kurtosis = 3, excess kurtosis = 0

3. Platykurtic 低峰态

1. Less peaked, thinner tailed than normal distribution.
2. Kurtosis < 3, excess kurtosis > 0

Time value of the money

1. Interest rate(r)

1. Required of return
2. Discount rate
3. Opportunity cost

2. Components of interest rate

1. Real risk free interest rate
2. Inflation premium
3. Risk premium

1. default risk premium
2. liquidity premium
3. maturity premium

4. Nominal interest rate

= Real risk-free interest rate + inflation premium + (default risk premium + liquidity premium + maturity premium)

5. Nominal risk-free interest rate = Real risk-free interest rate + inflation premium

3. Effective annual rate

1. $EAR = \left(1 + \frac{r_s}{m}\right)^m - 1$

4. Continuous compounding

1. $EAR = e^{r_s} - 1$

5. Annuity 年金

1. Ordinary annuity 普通年金:
2. Annuity due: 期初应付年金
 1. 金融计算器BGN模式