Backtesting overfitting

Barry Quinn

Table of contents

0.1 Outline

- Backtesting and selection bias under multiple testing
- Precision and recall in statistics
- Neyman Pearson Type I and Type || errors under multiple hypothesis testing

- False discovery
- Most important simulation in quantitative finance

0.2 Experiment evidence using simulation

- So far we have used experimental evidence extensively.
- More precisely we have used monte carlo simulations to allow us to reach conclusions regarding the mathematical properties of various estimators and algorithms under controlled conditions.
- Good financial research requires the ability to control for the conditions of an experiment that can result in *realistic* causal inference statements.

0.3 What is a backtest?

- A backtest is a historical simulation of how an investment strategy would have performed in the past.
- It is not a controlled experiment, because we cannot change the environmental variables to derive a new historical time series on which to perform an independent backtest.
- As a result, backtests cannot help us derive the precise cause–effect mechanisms that make a strategy successful.
- This identification issue is more than a techical inconvenience

0.4 Overfitting and statistical inflation

- In the context of strategy development, all we have is a few (relatively short, serially correlated, multicollinear and possibly nonstationary) historical time series.
- It is easy for a researcher to overfit a backtest, by conducting multiple historical simulations, and selecting the best performing strategy (Bailey et al. 2014).
- When a researcher presents an overfit backtest as the outcome of a single trial, the simulated performance is inflated.
- This form of statistical inflation is called selection bias under multiple testing (SBuMT).
- SBuMT leads to false discoveries: strategies that are replicable in backtests, but fail when implemented.

0.5 Backtest hyperfitting

- SBuMT is compounded as a consequence of sequential SBuMT at two levels:

1. Each researcher runs millions of simulations, and presents the best (overfit) ones to her boss
2. The company further selects a few backtests among the (already overfit) backtests submitted by the researchers.

- We may call this backtest hyperfitting, to differentiate it from backtest overfitting (which occurs at the researcher level).
- It may take many decades to collect the future (out-of-sample) information needed to debunk a false discovery that resulted from SBuMT.
- In this lecture we study how researchers can estimate the effect that SBuMT has on their findings.

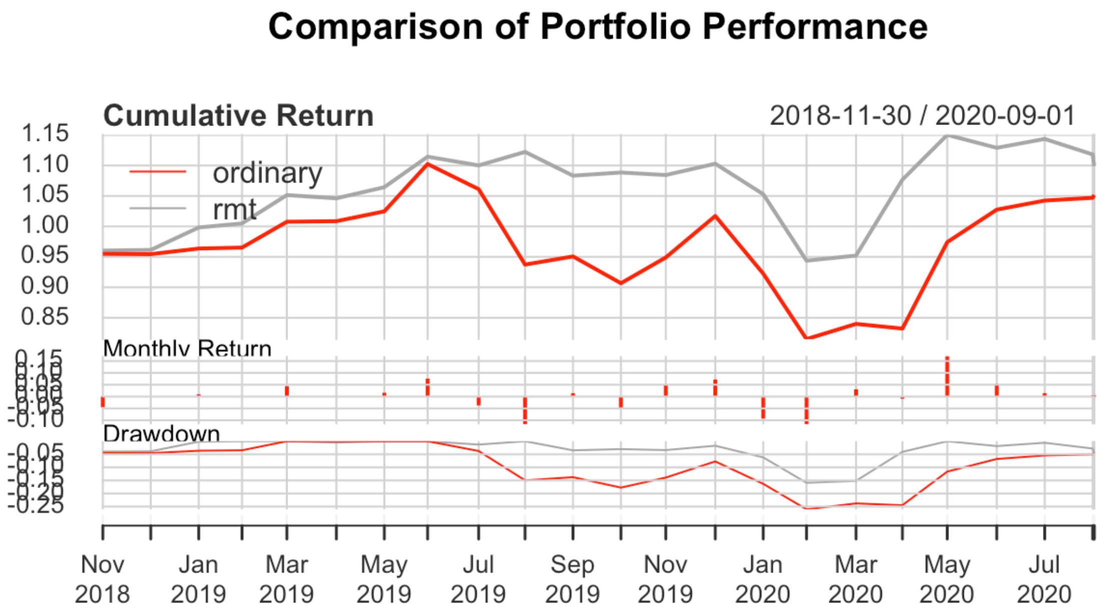
0.6 Performance statistics

Performance Statistic	Description
PnL	The total amount of dollars (or the equivalent in the currency of denomination) generated over the entirety of the backtest, including liquidation costs from the terminal position.
PnL from long positions	The portion of the PnL dollars that was generated exclusively by long positions.

Performance Statistic	Description
Annualized rate of return	The time-weighted average annual rate of total return, including dividends, coupons, costs, etc.
Hit ratio	The fraction of bets that resulted in a positive PnL.
Average return from hits	The average return from bets that generated a profit.
Average return from misses	The average return from bets that generated a loss.

0.7 Risk statistics

- Intuitively, a drawdown (DD)is the maximum loss suffered by an investment between two consecutive high-watermarks (HWMs).
- The time under water (TuW) is the time elapsed between an HWM and the moment the PnL exceeds the previous maximum PnL.
- In workshop 4 we used `PortfolioAnalytics` and chart the performance of our competing strategies.



- You can see the drawdown statistics in the bottom graph

0.8 Implementation shortfall statistics

0.8.0.1 Broker fees per turnover

- Broker fees per turnover: These are the fees paid to the broker for turning the portfolio over, including exchange fees.

0.8.0.2 Average slippage per turnover

- Average slippage per turnover: These are execution costs, excluding broker fees, involved in one portfolio turnover.

0.8.0.3 Dollar performance per turnover

- Dollar performance per turnover: This is the ratio between dollar performance (including brokerage fees and slippage costs) and total portfolio turnovers.

0.8.0.4 Return on execution costs

- Return on execution costs: This is the ratio between dollar performance (including brokerage fees and slippage costs) and total execution costs.

0.9 Efficiency statistics

Efficiency statistics provide a relative analysis of the performance of a backtest.

0.9.0.1 Annualized Sharpe ratio

- Annualized Sharpe ratio: This is the SR value, annualized by a multiplying by $\sqrt{a}$ (a=average number of returns observations per year).

0.9.1 Information ratio

- Information ratio: This is the SR equivalent of a portfolio that measures its performance relative to a benchmark.

0.9.2 Probabilistic Sharpe ratio

- Probabilistic Sharpe ratio: PSR corrects SR for inflationary effects caused by non-Normal returns or track record length.

0.9.3 Deflated Sharpe ratio

- Deflated Sharpe ratio: DSR corrects SR for inflationary effects caused by non-Normal returns, track record length, and selection bias under multiple testing.

0.10 Precision and Recall in Statistics

- To understand how false discoveries affect performance in algorithmic trading and investment, we must first introduce two concepts.
- In machine learning statistics, precision and recall are measures of task specific accuracy, especially in classification problems.
- In terms of investment strategies:

precision is the estimated probability that a randomly selected investment strategy from the pool of all positive backtests is a true strategy.

recall (or true positive rate) is the estimated probability that a strategy randomly selected from the pool of true strategy has a positive backtest

0.11 The Neyman-Pearson Framework

Under the standard Neyman-Pearson [1933] hypothesis testing framework:

- We state a null hypothesis H0, and an alternative hypothesis H1
- We derive the distribution of a test statistic under H0 and under H1
- We reject H0 with confidence $1 - \alpha$ in favour of H1 when we observe an event that, should H0 be true, should only occur with probability α

- This framework is the statistical analogue to a **proof by contradiction** argument
- There are 4 probabilities associated with a predicted positive $x > \tau_\alpha$

- $Pr(x > \tau_\alpha | H_0) = \alpha$ the type I error probability, or significance or false positive rate
- $Pr(x > \tau_\alpha | H_1) = 1 - \beta$ is the power of the test, recall or true positive rate, $Pr(x \leq \tau_\alpha | H_1) = \beta$ is the type II error probability or false negative rate
- $Pr(H_0 | x > \tau_\alpha)$ the false discovery rate (FDR)
- $Pr(H_1 | x > \tau_\alpha)$ the test's precision

- Note again that p-value α does not give the probability that the null hypothesis is true.

0.12 A mathematical argument (Lopez de Prado 2020)

- Let's say you have s investment strategies to analyze as a quant researcher.
- Inevitably, some of these strategies are false discoveries, in the sense that their expected return is not positive.
- Mathematically, we can denote:

$$s = s_T + s_F \text{vvwhere } s_T = \text{number of true strategies} s_F = \text{number of false strategies}$$

- Let θ be the odds ratio of true strategies against false strategies, $\theta = s_T / s_F$.

0.13 A mathematical argument (Lopez de Prado 2020)

- In finance, where the signal-to-noise ratio is low, false strategies abound, hence θ is expected to be low. The number of true investment strategies is:

$$S_T = s \times \frac{s_T}{s_T + s_F}$$

- Likewise, the number of false investment strategies is:

$$S_F = S - S_T = s\left(1 - \frac{\theta}{(1+\theta)}\right) = s\frac{1}{(1+\theta)}$$

- Given a false positive rate α (type I error), we will obtain a number of false positives, $FP = \alpha \times S_F$, and a number of true negatives, $TN = (1-\alpha)s_F$.

0.14 A mathematical argument (Lopez de Prado 2020)

- Denote β the false negative rate (type II error) associated with that α.
- We will obtain a number of false negatives, $FN = \beta \times s_F$, and a number of true positives, $TP = (1-\beta)s_T$.
- Thus:

$$\text{precision} = \frac{TP}{(TP+FP)} = \frac{(1-\beta)s_T}{(1+\beta)s_T + \alpha s_F} = \frac{(1-\beta)s\frac{\theta}{(1+\theta)}}{(1-\beta)s\frac{\theta}{(1+\theta)} + \alpha s\frac{\theta}{(1+\theta)}} = \frac{(1-\beta)\theta}{(1-\beta)\theta + \alpha}$$

$$\text{recall} = \frac{TP}{(TP+FN)} = \frac{(1-\beta)s_T}{(1-\beta)s_T + \beta s_T} = 1 - \beta$$

0.15 A mathematical argument (Lopez de Prado 2020)

- What the mathematical logic tells us is before running backtests on a strategy, researchers should gather evidence that a strategy may indeed exist.
- The reason is, the precision of the test is a function of the odds ratio θ.
- If the odds ratio is low, the precision will be low, even if we get a positive with high confidence (low p-value).

This is evidence to the pitfall that p-values report a rather uninformative probability. It is possible for a statistical test to have high confidence (low p-value) and low precision.

In particular, a strategy is more likely false than true if $(1-\beta)\theta < \alpha$ such that precision is less than 50%.

- Finally, there is an important relationship between the false discovery rate (FDR) and precision.
- Specifically,

$$FDR = \frac{FP}{(FP+TP)} = \frac{\alpha}{(1-\beta)\theta + \alpha} = 1 - precision$$

0.16 A FDR function

- The following is a simple function which calculates precision, recall and the false discovery rate.

```
fdr_anal <- function(ground_truth, alpha = 0.05, beta, trails) {
  theta = ground_truth / (1 - ground_truth)
  recall = 1 - beta
  b1 = recall * theta
  precision = b1 / (b1 + alpha)
  tibble(Recall = recall, Precision = precision, FDR = 1 - precision)
}
```

- Suppose before running backtests on a strategy, the researcher knows the *truth* that there is a 1% chance that the strategy is profitable.
- If she sticks with the standard convention of 5% significance level and a 20% chance of a false negative, and runs 1000 trails, what is the rate of false discoveries?

```
fdr_anal(0.01, beta = 0.2)
```

```
# A tibble: 1 x 3
  Recall Precision   FDR
   <dbl>     <dbl> <dbl>
1    0.8     0.139 0.861
```

- For this reason alone, we should expect that most discoveries in financial econometrics are likely false.

0.17 Familywise Error Rate (FWER)

- When Neyman and Pearson [1933] proposed this framework, they did not consider the possibility of conducting multiple tests and select the best outcome.
- When a test is repeated multiple times, the combined α increases.
- Consider that we repeat for a second time a test with false positive probability α.
- At each trial, the probability of not making a Type I error is $1 - \alpha$
- If the two trials are independent, the probability of not making a Type I error on the first and second tests is $(1 - \alpha)^2$
- The probability of making *at least one* Type I error is the complementary, $1 - (1 - \alpha)^2$
- After a *family* of K independent tests, we reject H0 with confidence $(1 - \alpha)^K$
- FWER the probability that at least one of the positives is false, $\alpha_K = 1 - (1 - \alpha)^K$
- The Sidak Correction: for a given K and α_K then $\alpha = 1 - (1 - \alpha_K)^{1/K}$

0.18 FWER vs FDR

- Thus far we have defined 2 Type 1 errors for multiple testing:

1. Familywise Error Rate (FWER): The probability that at least one false positive takes place.
2. False Discovery Rate (FDR): Expected value of the ratio of false positives to predicted positives.

- In most scientific and industrial applications, FWER is considered overly punitive.
  - For example, it would be impractical to design a car model where we control for the probability that a single unit will be defective.

0.19 FWER vs FDR

- However, in the context of finance, the FDR is preferrred as an investor does not typically allocate funds to all strategies with predicted positives within a family of trials, where a proportion of them are likely to be false.

- Instead, investors are only introduced to the single best strategy out of a family of thousands or even millions of alternatives
- Investors have no ability to invest in the discarded predicted positives.
- Following the car analogue, in finance there is actually a single car unit produced per model, which everyone will use. If the only produced unit is defective, everyone will crash.]

0.20 What does this all mean for quantitative finance

- Selection bias under multiple backtesting makes it impossible to assess the probability that a strategy is false.
- Lopez de Prado (2018) argues that this explains why most quantitative investment firms fail as they are likely investing in false positives
- This is because most financial analysts typically assess performance on the Sharpe ratio, not precision and recall.
- Lopez de Prado (2020) develops a framework to assess the probability that a strategy is false, using the Sharpe ratio estimate and metadata from the discovery process as inputs

0.21 The golden age of the Sharpe Ratio (1966-2012)

- In 1966, William Sharpe proposed a ratio metric that would come to dominate investment strategy appraisal
- Consider an investment strategy with excess returns (or risk premia) $r_t, t = 1, ..., T$ which follows an IID Normal distribution
$$r_t \sim N(\mu, \sigma)$$
- Non-annualised SR of such a strategy is defined as
$$SR = \frac{\mu}{\sigma}$$
- as the parameters μ and σ are unknown, they must be estimated such that SR is estimated as:
$$\hat{SR} = \frac{E(r_t)}{\sqrt{V_{r_t}}}$$

0.22 2002 Andrew Lo and Elmar Mertens

- Andrew Lo show that under the assumption that $r_t \overset{IID}{\sim} N(\mu, \sigma)$ the asymptotic distribution of $\hat{SR}$ is
$$(\hat{SR} - SR) \overset{a}{\to} N\left[0, \frac{1 + 0.5SR^2}{T}\right]$$
- Subsequent evidence showed hedge fund returns exhibit substantial negative skewness, and positive excess kurtosis.
- the implication being that assumed IID normal returns will grossly underestimate the false positive probability
- Elmar Mertons then derived an asymptotic distribution for $\hat{SR}$ that include a variance terms which incorporated skewness and kurtosis.

0.23 2012 David Bailey and Marco lopez de Prado

- In the Journal of Risk, David Bailey and Marco Lopez de Prado utilises previous results to derive the Probabilistic Sharpe Ratio
- PSR estimates the probability that the observed $\hat{SR}$ exceeds SR* as:

$$P\hat{S}R(SR*) = Z\left(\frac{(\hat{SR} - SR*)\sqrt{T-1}}{\sqrt{1-\hat{\gamma_3}\hat{SR}+\frac{\hat{\gamma_4}-1}{4}\hat{SR}^2}}\right)$$

- where Z[.] is the cumulative density function of the standard Normal distribution, T is number of observed returns, and $\hat{SR}$ is the non-annualised estimate of SR, computed on the same frequency as the T observations.

0.24 Inference on the Probabilistic Sharpe Ratio

- For a given SR*, the probabilistic sharpe ratio increases with greater mean returns, lower variance of returns, longer track record (T), positively skewed returns, and thinner tails

0.25 The False Strategy Theorem

- Bailey et al. (2014) formalised a theorem ,*False Strategy Theorem* , that expressed the SBuMT as a function on the number of trails and the variance of the Sharpe ratios.

- In practice a researcher may carry out a large number of historical simulations (trails) and report only the best outcome (maximum Sharpe ratio)
- Maximum Sharpe ratio is not randomly distributed which gives rise to *SBuMT*, so when more than one trail takes place the maximum Sharpe ration is greater than the expected value of the Sharpe ration from a random trail.
- The theorem shows that given a investment strategy with an expected Sharpe ratio of zero and non-zero variance, the expected value of the maximum Sharpe ratio is strictly positive and a function of the number of trails

0.26 The False Strategy Theorem

- Given a sample of IID-Gaussian Sharpe ratios $\widehat{SR_k}, k = 1, .., K$ with $\widehat{SR_k} \sim N(0, V(\widehat{SR_k})$

$$E(\underset{k}{max}(\widehat{SR_k}))V(\widehat{SR_k}^{-0.5} \approx (1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right]+\gamma Z^{-1}\left[1-\frac{1}{Ke}\right]$$

- where Z^{-1} is the inverse of the standard Gaussian CDF, e is Euler's number, and $ $is the Euler-Mascheroni constant.

- **Corollary:** Unless $\underset{k}{max}(\widehat{SR_k}) >> E(\underset{k}{max}(\widehat{SR_k}))$ the discovered strategy is likely to be a false positive.

- But $E(\underset{k}{max}(\widehat{SR_k}))$ is usually unknown, ergo SR is dead.

0.27 The *False Strategy* theorem

- .Lopez de Prado (2020) 🐍 code
- The theorem can be used to express the magnitude of the SBuMT as the difference between the expected maximum Sharpe ratio and the expected Sharpe ratio of a *false* strategy from a random trail

0.28 The *False Strategy* theorem R

```
getExpectedMaxSR<-function(nTrails,meanSR,stdSR){
  # Expected Max SR controlling for SBuMT
  emc=0.5772156649015328606065120900824024310421593336
  sr0=(1-emc)*qnorm(p=1-1./nTrails)+emc*qnorm(1-(nTrails*exp(1))^(-1))
  sr0=meanSR+stdSR*sr0
  return(sr0)
}
```

0.29 Distribution of Maximum SR

```
getDistMaxSR<-function(nSims,nTrails,meanSR,stdSR){
  out=tibble("Max{SR}"=NA,"nTrails"=NA)
  for (nTrails_ in nTrails) {
    #1) Simulated Sharpe Ratios
    set.seed(nTrails_)
    sr<-array(rnorm(nSims*nTrails_),dim = c(nSims,nTrails_))
    sr<-apply(sr,1,scale) # demean and scale
    sr= meanSR+sr*stdSR
    #2) Store output
    out<-out %>% bind_rows(
      tibble("Max{SR}"=apply(sr,2,max),"nTrails"=nTrails_))
  }
  return(out)
}
```

0.30 Run the experiment

```
library(pracma)
# Create a sequential on the log-linear scale
nTrails<-as.integer(logspace(1,4,100)) %>% unique()
plot(nTrails)
sr0=array(dim = length(nTrails))
for (i in seq_along(nTrails)) {
  sr0[i]<-getExpectedMaxSR(nTrails[i],meanSR = 0, stdSR = 1)
}
sr1=getDistMaxSR(nSims = 1000,nTrails = nTrails,meanSR = 0,stdSR = 1)
```

0.31 Most important plot in Quantitative finance

0.32 Inference from plot

- The experiment compares the empirical (Monte Carlo) estimate of Maximum Sharpe ratio under the null of a false strategy to that implied by the FS theorem
- The plot shows the output of the experiment for 1 to 10,000 trails.
- The code sets $V[\hat{SR}_k] = 1$ and simulates the maximum Sharpe ratio 500 times, to derive a distribution of maximum Sharpe ratios for any k (number of trails).
- the y axis shows the distribution of the $max_k(\hat{SR}_k)$ and the Expect

- this results is profound, after only 100 independent backtests the expected maximum Sharpe ratio is 3.2, even when the true Sharpe ratio is zero.
- The reason is **Backtest overfitting**: when selection bias (picking the best results) takes place under multiple testing (running many alternative configurations) that backtests are likely to be false discoveries.

0.33 A Solution

- The Deflated Sharpe Ratio computes the probability that the Sharpe Ratio (SR) is statistically significant.

$$\widehat{DSR} \equiv \widehat{PSR}(\widehat{SR_0}) = Z\left[\frac{(\hat{SR} - E[max_k(\widehat{SR_k})])\sqrt{T-1}}{\sqrt{1-\hat{\gamma_3}\widehat{SR}+\frac{\hat{\gamma_4}-1}{4}\widehat{SR}^2}}\right]$$

- $\widehat{DSR}$ can be interpreted as the probability of observing a Sharpe ratio greater or equal to $\widehat{SR}$ subject to the null hypothesis that the true Sharpe ratio is zero, while adjusting for skewness γ_3, kurtosis γ_4, sample length and multiple testings.

- Calculate DSR requires the estimation $E[max_k(\widehat{SR_k})])$ which requires estimating K and $V(\hat{SR})$ which is where FML can help.

- Specifically, we are employ optimal number of clustering to estimate K the effective number of trails and then calculate the variances.

0.34 Implications for Academics

- Most studies in empirical finance are false (Harvey et al., 2016)
- Selection bias may invalidate the entire body of work performed for the past 100 years
- Finance cannot survive as a discipline unless we solve this problem
- Investors and regulators have no reason to trust the value added by researchers and asset managers unless we learn to prevent false discoveries

- Applying the False Strategy theorem to prevent false positives in finance
- Requires estimating two meta-research variables to discount for "lucky findings"
- Academic journals should cease accepting papers that do not control for selection bias under multiple testing
- Papers must report the probability that the claimed financial discovery is a false positive

0.35 Implications for Regulators

- Before the FDA, adulteration and mislabeling of food and drugs caused frequent episodes of mass poisoning, birth defects, and death
- Financial firms engaging in backtest overfitting defraud investors for tens of billions of dollars annually
- The SEC could demand quantitative firms certify the probability that promoted investments are bogus
- Quantitative firms should be required to store all trials involved in a discovery for post-mortem analysis

0.36 Implications for Investors

- Many financial firms promote pseudo-scientific products as scientific
- Investment products based on award-winning journal articles are not necessarily scientific
- If the original author has not become rich with the discovery, investors' chances are slim
- Investors should demand firms report the results of all trials, not only the best-looking ones
- Investors should consult databases of investment forecasts and assess the credibility of gurus and financial firms based on all outcomes from past predictions

0.37 References

- Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical Asset Pricing via Machine Learning." The Review of Financial Studies.
- Harvey, Campbell R., Presidential Address: The Scientific Outlook in Financial Economics. 2017.
- American Statistical Association. 2016. "Ethical guidelines for statistical practice."
- López de Prado, M. and M. Lewis. 2018. "Detection of False Investment Strategies Using Unsupervised Learning Methods."
- Bailey, D., J. Borwein, M. López de Prado, and J. Zhu. 2014. "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance."
- Bailey, D., J. Borwein, M. López de Prado, and J. Zhu. 2017. "The Probability of Backtest Overfitting."
- Bailey, D. and M. López de Prado. 2012. "The Sharpe ratio efficient frontier."
- Bailey, D. and M. López de Prado. 2014. "The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality."