Lecture 2-b
Introduction: Review, Returns & Data

(for private use, not to be posted/shared online)

1

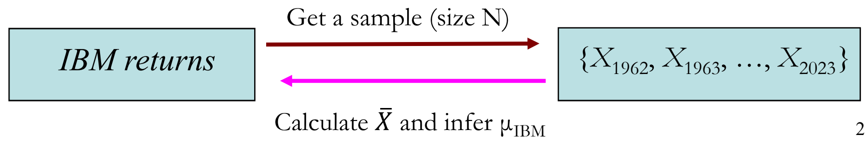
Review – Sample Statistics

• Sample Statistics: A function of data. Used to estimate population parameters

Example: $\bar{X}$ is an estimate of the population mean, μ.

Notation: Population parameters: Greek letters (μ, σ, θ, etc.).
Estimators: A hat over the Greek letter ($\hat{\theta}$).

Suppose we want to learn about annual mean IBM returns, μ_{IBM}.
1. Get a sample: $\{X_{1962}, X_{1963}, \ldots, X_{N=2023}\}$.
2. Compute a statistic, $\bar{X}$.



2

Review – Sample Statistic

• Some simple properties for $\hat{\theta}$:

- $\hat{\theta}$ is *unbiased* estimator of θ if $E[\hat{\theta}] = \theta$.

- $\hat{\theta}$ is *most efficient* if the variance of the estimator is minimized.

- $\hat{\theta}$ is Best Unbiased Estimate (BUE), if it is the estimator with the smallest variance among all unbiased estimates.

- $\hat{\theta}$ is *consistent* if as the sample size, *n*, increases to ∞, $\hat{\theta}_n$ converges to θ. We write: $\hat{\theta}_n \xrightarrow{p} \theta$.

- $\hat{\theta}$ is *asymptotically normal* if as the sample size, *n*, increases to ∞, $\hat{\theta}_n$, often standardized or transformed, converges in distribution to a Normal distribution. We write $\hat{\theta}_n \xrightarrow{d} N(\theta, Var(\hat{\theta}_n))$.

3

Review – PDF for a Continuous RV

Definition: Suppose that *X* is a random variable. Let $f(x)$ denote a function defined for $-\infty < x < \infty$ with the following properties:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$
3. $P[a \leq X \leq b] = \int_a^b f(x)dx$

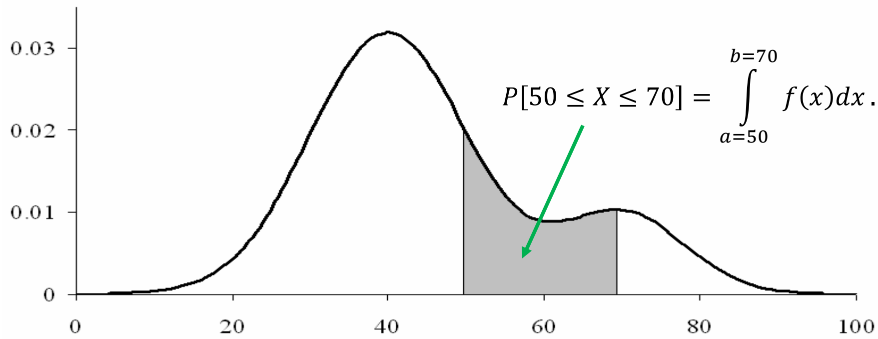
• Then, $f(x)$ is called the *probability density function* (pdf) of *X*. The RV *X* is called *continuous.* We use the pdf to describe the behavior of *X*.

• Analogous definition applies for a discrete RV, where the notation uses $p(x)$ instead.

4

Review – PDF for a Continuous RV

• The pdf is non-negative and integrates to $\int_{-\infty}^{\infty} f(x)dx = 1$.



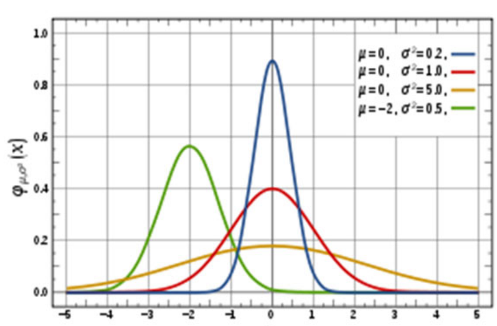
• Remark: We use the pdf to describe the behavior of the RV (discrete or continuous).

5

Review – Popular PDFs: Normal Distribution

• A RV *X* is said to have a *normal distribution* with parameters μ *(mean)* and σ^2 *(variance)* if *X* is a continuous RV with pdf $f(x)$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Note: Described by two parameters: μ and σ^2. We write $X \sim N(\mu, \sigma^2)$

6

Review – Popular PDFs: Gamma Distribution

- Let the continuous RV X have density function):

$$f(x) = \begin{cases} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\alpha, \lambda > 0$ and $\Gamma(\alpha)$ is the gamma function evaluated at α.

Then, X is said to have a *Gamma distribution* with parameters α and λ., denoted as $X \sim \text{Gamma}(\alpha, \lambda)$ or $\Gamma(\alpha, \lambda)$.

It is a family of distributions, with special cases:

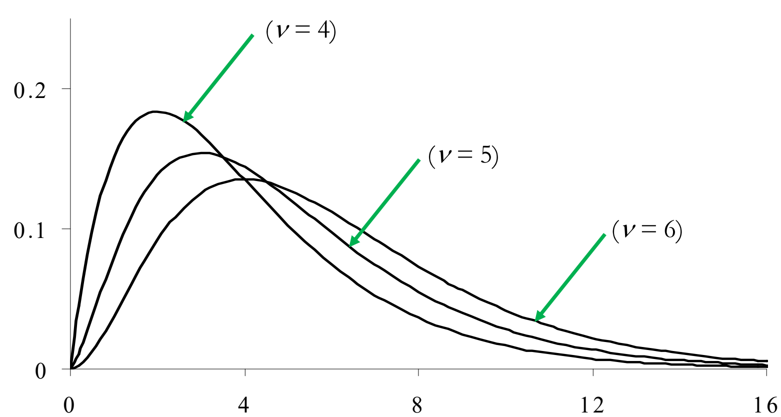
– Exponential Distribution, or Exp(λ): $\alpha = 1$.

– **Chi-square Distribution**, or χ^2_ν: $\alpha = \nu/2$ and $\lambda = ½$.

7

Review – Popular PDFs: Gamma Distribution

- Below we plot the Chi-square distribution with parameter ν, which we refer as degrees of freedom:



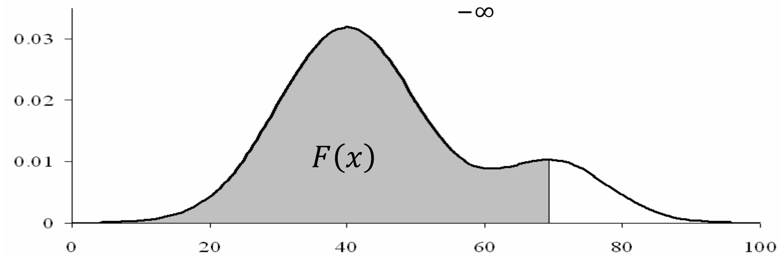
Note: When ν is large, the χ^2_ν converges to a N(ν, 2ν).

8

Review – CDF for a Continuous RV

• If X is a continuous random variable with probability density function, $f(x)$, the *cumulative distribution function* (CDF) of X is given by:

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)dt\,.$$



• Note: The FTC (*fundamental theorem of calculus*) implies:

$$F'(x) = \frac{dF(x)}{dx} = f(x)$$

9

Review – The Empirical Distribution

• The empirical distribution (ED) of a dataset is simply the distribution that we observe in the data.

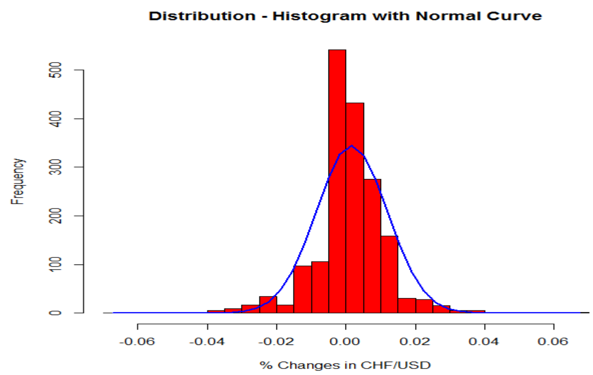
The ED is a discrete distribution that gives equal weight to each data point, assigning a $1/N$ probability to each of the original N observations.

We often use a histogram to visualize the ED.

10

Review – Histogram of a RV

Example: We use a histogram to estimate the distribution of a RV. Let X = Percentage changes in the CHF/USD exchange rate = e_f
Data: Monthly from January 1971 to June 2020 (N=594 observations).



Note: We overlay a Normal density (blue line) over the histogram. 11

Review – Moments of Random Variables

• The moments of a random variable X are used to describe the behavior of the RV (discrete or continuous).

Definition: K^{th} Moment
Let X be a RV (discrete or continuous), then the *k^{th} moment of* X is:

$$\mu_k = E(X^k) = \begin{cases} \sum_x x^k p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

• The first moment of X, $\mu = \mu_1 = E(X)$ is the center of gravity of the distribution of X.

• The higher moments give different information regarding the shape of the distribution of X. 12

Review – Moments of a RV

Definition: Central Moments

Let X be a RV (discrete or continuous). Then, the k^{th} *central moment* of X is defined to be:

$$\mu_k^0 = E[(X-\mu)^k] = \begin{cases} \sum_x (x-\mu)^k p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

where $\mu = \mu_1 = E(X)$ = the first moment of X.

• The central moments describe how the probability distribution is distributed about the center of gravity, μ. First two:

$\mu_1^0 = E[X-\mu] = 0$

$\mu_2^0 = E[(X-\mu)^2] = \text{Var}[X] = \sigma^2$

13

Review – Estimating Moments

• We estimate expected values with sample averages. For example, the first moment, μ, & the second central moment, σ^2, are estimated by:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$s^2 = \frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{N-1}$$ (N-1 adjustment needed for $E[s^2] = \sigma^2$)

• Besides consistent, they are both are *unbiased* estimators of their respective population moments (unbiased = "on average, I get the population parameter"). That is,

$E[\bar{X}] = \mu$

$E[s^2] = \sigma^2$ "μ & σ^2 population parameter"

14

Review – LLN & CLT

- Under certain assumptions (finite mean & variance) for X_1, X_2, …, X_N, a sequence of *i.i.d.* RVs, we have:

LLN

A sample average as the sample size goes to infinite tends to its expected value. Also written as $\bar{X}_N \xrightarrow{p} \mu$. *(convergence in probability)*

CLT

The sample mean approaches the normal distribution with mean μ and variance σ^2/N. This theorem is sometimes stated as

$$\frac{\sqrt{N}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1)$$ (*convergence in distribution*).

15

Sampling Distributions: Mean

- All statistics, T(*X*), are functions of RVs and, thus, they have a distribution. The finite sample distribution of T(*X*) is called the *sampling distribution*.

- For the sample mean $\bar{X}$, if the X_i's are normally distributed, then the sampling distribution is normal with mean μ and variance σ^2/N. Or

$$\bar{X} \sim N(\mu, \sigma^2/N).$$

Then, $E[\bar{X}] = \mu$

$Var[\bar{X}] = \sigma^2/N$ ⇒ $Var[\bar{X}]$ decreases as N increases!

- The SD of the sampling distribution is called the *standard error* (SE). Then, $SE(\bar{X}) = \sigma/\text{sqrt}(N)$.

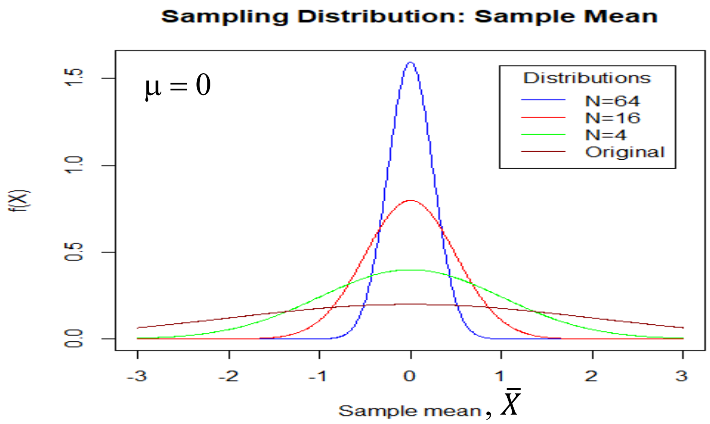
- We associate the SE with the precision of the estimate. The precision of the estimation of the mean increases as N increases.

16

Sampling Distributions: $\bar{X}$

- Sampling Distribution for the Sample Mean of a normal population:

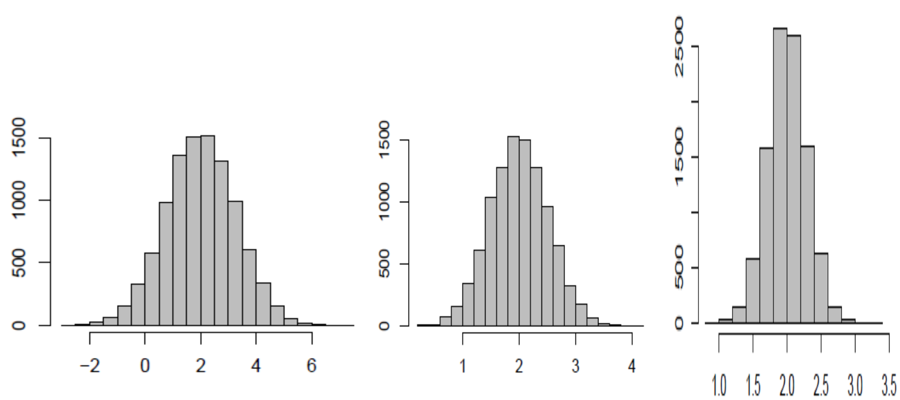
$$\bar{X} \sim N(\mu, \sigma^2/N)$$



Note: As $N \rightarrow \infty$, $\bar{X} \rightarrow \mu$ –i.e., the distribution becomes a spike at μ![17]

Sampling Distributions: $\bar{X}$

- 10,000 samples, for a N(2, 4) population. Different sample sizes, N:



N = 10 N = 60 N = 200

Note: As $N \rightarrow \infty$, $\bar{X}$ becomes more concentrated around μ = 2. In the limit, a spike at μ! 18

Sampling Distributions: Variance

- For the sample variance, s^2, if the X_i's are normally distributed, the sampling distribution is derived from this result:

$$\frac{(N-1)\, s^2}{\sigma^2} \sim \chi^2_{N-1}.$$

Then, $E[s^2] = \sigma^2$

$Var[s^2] = 2 * \sigma^4 / (N-1)$

$SE(s^2) = \frac{\sqrt{2}\ \sigma^2}{\sqrt{N-1}}$ $\Rightarrow$ precision increases as N increases!

Remark: The precision of the estimation increases as N increases.

This remark is especially relevant in Finance, where we derive relations between expected returns and risk factors, like market risk or volatility. As we gather more data, expected returns and the volatility of returns will be more precisely estimated.

19

Review – Sampling Distributions: s^2

- Summary for s^2 of normal variates:

Sampling distribution: $(N-1)\, s^2/\sigma^2 \sim \chi^2_{N-1}$.

Mean: $E[s^2] = \sigma^2$

Variance: $Var[s^2] = 2 * \sigma^4/(N-1)$.

- If the data is **not normal** (& ***N*** is **large**), the CLT can be used to approximate the sampling distribution by the asymptotic one:

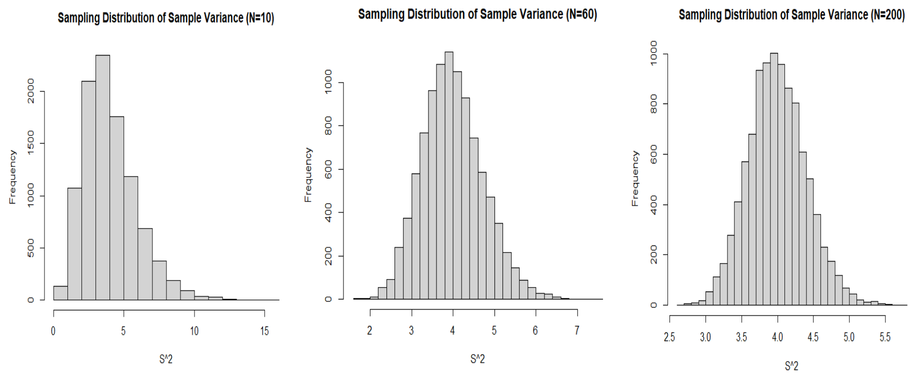
$$s^2 \xrightarrow{a} N(\sigma^2, \sigma^4 * (\kappa - 1)/\, N)$$

where $\kappa = \frac{\mu_4^0}{\sigma^4}$ (recall when data is normal, $\kappa = 3$).

20

Sampling Distributions: s^2

• 10,000 samples, for a N(2, 4) population. Different sample sizes, *N*:



$N = 10$ $N = 60$ $N = 200$

Note: As $N \to \infty$, the distribution of s^2 looks more Normal – the CLT at work!

21

Returns

- Net or simple (total) return, R_t:

$$R_t = \frac{(P_t - P_{t-1}) + D_t}{P_{t-1}} = \text{capital gain + dividend yield}$$

where P_t = Stock price or Value of investment at time t
D_t = Dividend or payout of investment at time t

Note: This is the return from time t-1 to time t. To be very explicit we can write this as $R_{t-1,t}$.

- Gross (total) Return: $R_t + 1 = \frac{P_t + D_t}{P_{t-1}}$

- In general, when the word "total return" is used in the definition, it means "*returns including dividends.*"

22

Returns – Log Returns

• There is another commonly used definition of return, the *log (total) return*, r_t, defined as the log of the gross return:

$$r_t = \log(1 + R_t) = \log(P_t + D_t) - \log(P_{t-1})$$

Note: When the values are small (-0.1 to +0.1), the two returns are *approximately* the same: $r_t = \log(1 + R_t) \approx R_t$.

In general –i.e., when returns are not small, $r_t < R_t$.

• When using daily and weekly data, both definitions are very similar. As the time interval of measurement of returns increases, the difference between the two measures also differs.

• In this class, we will use log returns.

23

Introduction to R

• We go over basic R commands:

- Use R as a calculator
- Assign values to variables <-
- Create data using c(), seq, rep, rnorm and runif
- Create functions
- Loops: "for" loop
- Graphs for data: plots and histograms
- Formulas for moments
- Application to exchange rates (CHF/USD): Reading data from a file, define log returns, plot data, and compute moments.

24

Estimating Moments in R

• We estimate sample averages for e_f = **% changes in the CHF/USD**, using data from 1971:Jan – 2020:Jun.

First, we need to import the data. In R, we use the **read** function, usually followed by the type of data we are importing. Below, we import a comma separated values (csv) file with monthly CPIs and exchange rates for 20 different countries, then we use the **read.csv** function:

```
PPP_da <-
read.csv("http://www.bauer.uh.edu/rsusmel/4397/ppp_2020_m.csv",head
=TRUE,sep=",")
```

Second, we extract from the imported data, PPP_da, the column corresponding to the CHF/USD exchange rate:

```
x_chf <- PPP_da$CHF_USD            # Extract CHF/USD exchange rate
```

25

Estimating Moments in R

• Now, we estimate the sample moments for e_f = log returns ≈ % changes in the CHF/USD.

```
T <- length(x_chf)                 # Size of data (T or N)
e_chf <- log(x_chf[-1]/x_chf[-T])  # log changes in CHF/USD FX rate

x <- e_chf                         # Series to be analyzed
n <- length(x)                     # Number of observations (N=593)
m1 <- sum(x)/n                     # Mean
m2 <- sum((x-m1)^2)/n              # Used in denominator of both
m3 <- sum((x-m1)^3)/n              # For numerator of S
m4 <- sum((x-m1)^4)/n              # For numerator of K
b1 <- m3/m2^(3/2)                  # Sample Skewness
b2 <- (m4/m2^2)                    # Sample Kurtosis
s2 <- sum((x-m1)^2)/(n-1)          # Sample Variance
sd_s <- sqrt(s2)                   # Sample SD
```

26

Estimating Moments in R

- Output:

```
> m1                          # Sample mean
[1] -0.002550636
> s2                          # Sample Variance
[1] 0.001115257
> sd_s                        # Sample SD
[1] 0.03339546
> b1                          # Sample Skewness
[1] -0.06733514
> b2                          # Sample Kurtosis
[1] 4.621602
> median(x)                   # Sample Median
-0.001431
```

- Small mean (**0.25%**), slight negative skewness, kurtosis greater than 3, pointing out to non-normality of data ("*fatter tails*").

27

Estimating Moments

Example: Summary of moments for e_f (log changes in CHF/USD):

Statistic	e_f
Mean	**-0.002551**
Median	-0.001431
Maximum	0.145542
Minimum	-0.145639
Std. Dev.	**0.033395**
Skewness	-0.067335
Kurtosis	4.621602

$\Rightarrow \gamma_2 = \frac{\mu_4^0}{\sigma^4} - 3 = 1.62$

Note: If the e_f's are normally distributed, then $\bar{X} \sim N(\mu, \sigma^2/N)$.
We estimate both moments:

Mean = **-0.002550636** (USD depreciated 0.25% vs CHF)

Variance[$\bar{X}$] = (**0.033395**)^2/593 = **0.880703e-06**

28

Multi-period Returns & Real Returns

- Multi-period holding return

k-period gross holding return $(1 + R_{t,t+k}) = \prod_{j=0}^{k-1}(1 + R_{t+j,t+j+1})$

With the log approximation: $r_{t,t+k} = \sum_{j=0}^{k-1} r_{t+j,t+j+1}$

- If $r_{t+j,t+j+1} = r \quad \Rightarrow \quad r_{t,t+k} = \sum_{j=0}^{k-1} r_{t+j,t+j+1} = k * r$

- If $r_{t+j,t+j+1}$ are independent with a constant variance equal to σ_r^2

$$\Rightarrow \text{Var}(r_{t,t+k}) = \sum_{j=0}^{k-1} var(r_{t+j,t+j+1}) = k * \sigma_r^2$$

$$\Rightarrow \text{SD}(r_{t,t+k}) = \sqrt{k * \sigma_r^2} = \sqrt{k} * \sigma_r$$

- Real returns $R_t^{real} = \frac{(1+R_t)}{(1+\pi_t)} - 1$

Log approximation $r_t^{real} \approx r_t - \pi_t$

29

Returns: Sample Moments

Example: Long-run S&P 500 monthly log returns & inflation, from Robert Shiller's website (1871:Jan -2021: Sep; N = **1805**).

```
Sh_da <-
read.csv("http://www.bauer.uh.edu/rsusmel/4397/Shiller_2021_m_data.csv",
head=TRUE, sep=",")

SP <- Sh_da$P                          # Extract SP = S&P Composite series
D <- Sh_da$D                           # Extract D = S&P Dividends series
CPI <- Sh_da$CPI                       # Extract CPI = Price Index series
R_SP <- Sh_da$Real_P                   # Extract R_SP = Real S&P500 series
i_10 <-  Sh_da$Long_i                  # Extract Long_i = Interest rates
T <- length(SP)

# Define log (percent) changes
lr <- log(SP[-1]/SP[-T])               # log capital gains
lr_t <- log(SP[-1] + D[-1]/12) - log(SP[-T])   # log (total) returns, includes dividends
I <- log(CPI[-1]/CPI[-T])              # Log Inflation rate
```

30

Returns: Sample Moments

Example: Long-run S&P 500 monthly log returns & inflation, from Robert Shiller's website (1871:Jan -2021: Sep; N = **1805**).

	Total Return	Capital Gains	Inflation	Real Total Return
Mean	**0.007378**	**0.003801**	**0.001707**	0.005615
Median	0.009928	0.006756	0.001432	0.009367
Maximum	0.414151	0.407459	0.068054	0.421504
Minimum	-0.30365	-0.30753	-0.06805	-0.30364
Std. Dev.	**0.040455**	**0.040650**	**0.010344**	**0.040726**
Skewness	**-0.47050**	-0.51512	-0.14404	-0.36439
Kurtosis	14.61046	14.35809	9.81997	14.33933
Jarque-Bera	**10205.0**	**9782.2**	**3504.3**	**9719.3**
P-value (JB)	2.2e-16	2.2e-16	2.2e-16	2.2e-16

31

Returns: Sample Moments

Example (continuation):

- Check some results from log approximation:

(1) $r_t^{real} \approx r_t - \pi_t$ (for small % changes)

0.005615 ≈ **0.007378** – **0.001707** = 0.005671

(2) Multi-period return: k = **12** –i.e., from monthly to annual,

- annual return = **0.007378** * **12** = **0.088531** (8.85%)

- annual SD = **0.040455** * sqrt(**12**) = 0.14014 (14.01%)

Note: Compounding the return: (1 + **0.007378**)^**12** – 1 = 0.0922. ¶

According to the annualized number, since 1871, the average total stock market return has been **8.85%** per year.

32

Returns: Sample Moments - Changing frequency

• Assuming independence of returns and constant moments, we use log returns to change frequencies for mean and variance of returns.

Suppose we have compounded data in base frequency ***b*** (say, monthly), but we are interested in compounded data in frequency *q* (say, annual). The approximation formulas for mean and standard deviation (SD) are:

• *q*-frequency mean = ***b***-freq mean * *q*/***b***

• *q*-freq SD = ***b***-freq SD * $\sqrt{}$(*q*/***b***) $\Rightarrow$ *q*-freq Variance = sqrt(*q*-freq SD)

Example: Using the monthly data from previous table we calculate the weekly mean and standard deviation for returns (***b*** = **30**, *q* = 7).

- weekly return = **0.007378** * (7/**30**) = 0.00172 (0.172%)

- weekly SD = **0.040455** * sqrt(7/**30**) = 0.01954 (1.95%)

Note: de-compounding the return: (1 + **0.007378**)^(7/**30**) = 0.00172.

33

Returns: Sampling Distribution

• Recall that the sampling distribution of the sample mean is:

$$\bar{X} \sim N(\mu, \sigma^2/N)$$

Example: Using Shiller's S&P 500 monthly log returns (N = **1805**).

Estimated Monthly mean return = $\bar{X}$ = **0.007378**

Estimated Var($\bar{X}$) = s^2/N = **0.040455**^2 /**1805** = 9.067075e-07

The SD of the monthly mean (also called the Standard Error, SE):

S.E.($\bar{X}$) = sqrt(9.067075e-07) = **0.000952** (or 0.095%).

$$\Rightarrow \bar{X} \sim N(0.007378, .00095^2)$$

Note: Compared to returns, expected returns, estimated by the sample mean, are more precisely estimated (**0.10%** vs **4.05%**). Not surprised, the sampling distribution of the mean shrinks towards the population mean as N increases.

34

Yields

- Consider an $n-$period discount bond. Time is measured in years. Today is t. Bond (asset) pays F_{t+n} dollars n years from now, at $t + n$.

F_{t+n} = Face value (value at time $t + n$).

P_t = Market price of the bond (Today's initial capital).

$r_{n,t}$ = Yield to maturity (YTM) at time t for a maturity of n years.

n = Maturity of bond.

$$P_t = \frac{F_{t+n}}{(1 + r_{n,t})^n}$$

Interpretation: If our initial capital, P_t dollars, is invested today at the interest rate $r_{n,t}$ for n years *compounded* annually, then, at time $t+n$, the payoff is F_{t+n}.

YTM, $r_{n,t}$, is a raw number like 4% at an annual rate. We use decimal notation, that is, is 0.04.

35

Continuous compounding

- More generally, suppose an investment is compounded m times per n years; where m is number of times return (yield) is compounded, for example, m = 4 for quarterly, m = 12 for monthly, $m = \infty$ for continuous compounding. Then, the market price of the bond is:

$$P_t = \frac{F_{t+n}}{(1+\frac{r_t}{m})^{mn}}$$

As the number of compounding periods grow, $m \rightarrow \infty$, the value of the payoff at time $t + n$ becomes:

$$F_{t+n} = P_t \, (1 + \frac{r_t}{m})^{mn} \rightarrow P_t \, e^{r_t n}$$

where we used $$\lim_{m\rightarrow\infty} \left(1 + \frac{x}{m}\right)^{mn} = e^{xn}$$

Then,

$$P_t = e^{-r_t n} \, F_{t+n}$$

36

Continuous compounding – Log Returns

- That is, with $m \rightarrow \infty$, we have :

$$P_t = e^{-r_t n} F_{t+n}$$

- Suppose the continuously compounded bond at maturity pays \$1 ($=F_{t+n}$) and the remaining duration is D units of time. Then,

$$P_t = e^{-r_t D} \$1 \qquad (\Rightarrow \log(P_t) = -D * r_t)$$

Then, the log return per unit of time is:

$$\log(P_{t+1}) - \log(P_t) = D * (r_t - r_{t+1})$$

where we ignore that D has one unit of time less at time t+1. That is, the daily returns of bonds is the change of yields multiplied by the duration.

37

Continuous compounding – Log Returns

- Now, suppose we invest P_0 in a bond with continuous compounding at an annual rate r. Then, the value of the investment at year t is:

$$V_t = P_0 e^{rt} \qquad (\Rightarrow \log(V_t) = \log(P_0) + t\, r)$$

The *log return* (log of gross return) *per year (unit of time)* is r:

$$\log(V_{t+1}) - \log(V_t) = [\log(P_0) + (t+1)\, r] - [\log(P_0) + t\, r] = r$$

The simple annual interest rate *r* quoted in the market is the annual log return if the interest is compounded continuously.

- The *effective annual interest rate,* r^a, is simple the annual rate of return:

$$(1 + r^a)^n = \left(1 + \frac{r}{m}\right)^{mn} \qquad \Rightarrow \qquad r^a = \left(1 + \frac{r}{m}\right)^m - 1$$

38

Bond Yields & Stock Returns

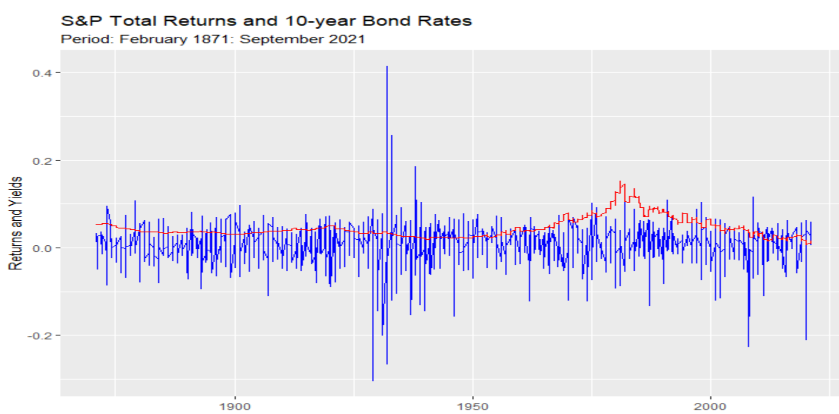
Example (continuation): We report the sample moments and descriptive statistics for **S&P total returns, 10-year interest rates** (annualized), and **total excess returns**, taken from Shiller's website.

	Total Return	10-year Interest Rates	Total Excess Returns
Mean	**0.007378**	**0.04511**	**0.003619**
Median	0.009928	0.0382	0.006395
Maximum	0.414151	0.1532	0.411264
Minimum	-0.30365	0.0062	-0.306435
Std. Dev.	**0.040455**	**0.02304**	**0.040519**
Skewness	-0.47050	1.79664	-0.450043
Kurtosis	14.61046	6.75173	14.51218
Jarque-Bera	**10204.9**	**2029.7**	**10028.0**
P-value (JB)	2.2e-16	2.2e-16	2.2e-16

39

Bond Yields & Stock Returns

Example (continuation): We plot long-run S&P returns (blue) and 10-year Bond interest rates (red) data.

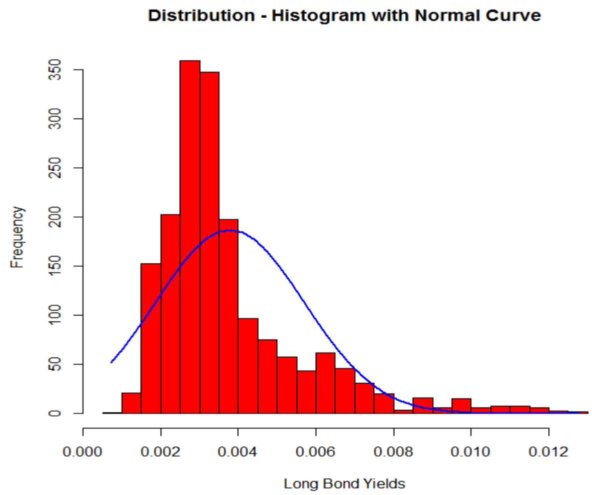


Note: Interest rates have a lower monthly mean (**0.003759** = **0.04511**/12), & lower volatility and kurtosis than stock returns.

40

Bond Yields: Distribution

Example (continuation): Bond interest rates data, from Robert Shiller's website:



41

