Hypothesis testing for the population variance with a chi-squared distribution

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

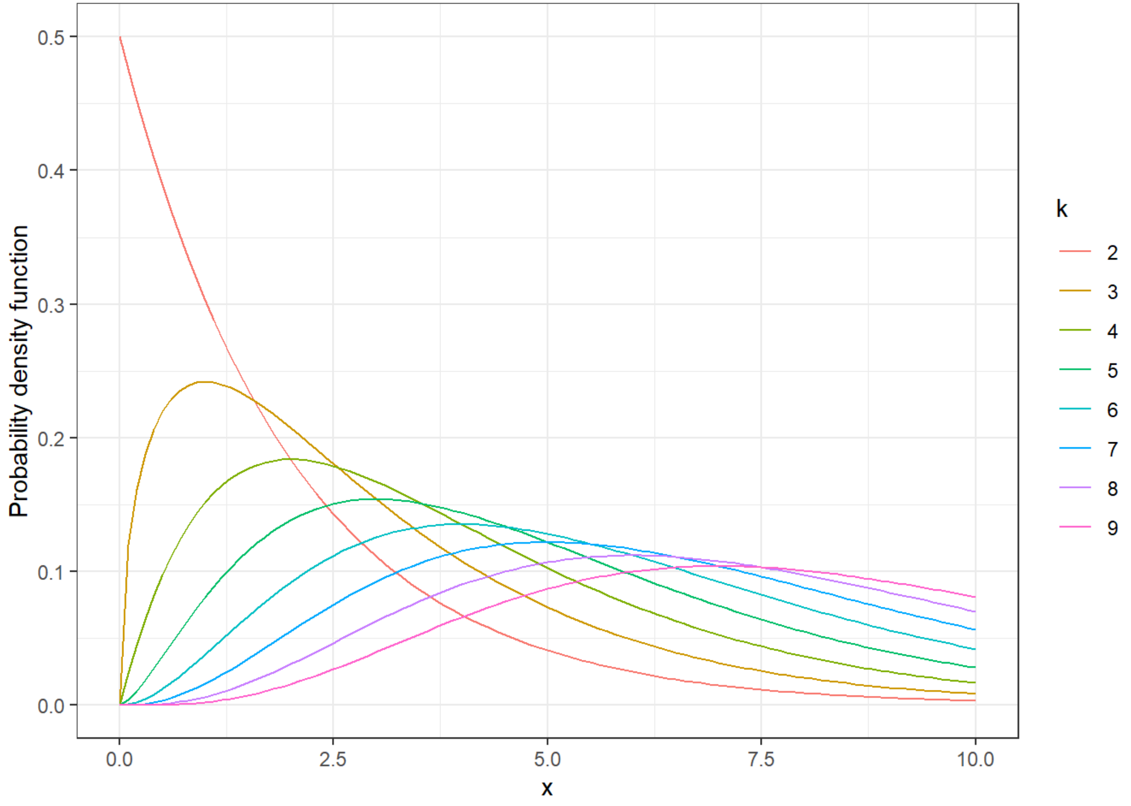
What will we cover today?

- We will look at the use of chi-squared distributions for hypothesis testing;
- We will look at an illustrative time series example our focus is the variance parameter;
- We will look at the distribution of a sample statistic involving the sample variance;
- We will use this distributional behavior to introduce the chi-squared test for population variance.

The chi-squared family of distributions

A random variable Q is said to be chi-squared with k degrees of freedom $Q \sim \chi^2(k)$ if

$$Q = \sum_{i=1}^{k} Z_i^2$$ with $Z_1, \cdots, Z_k \sim \mathcal{N}(0,1)$ independent and identically distributed.



Modelling a time series of stock prices

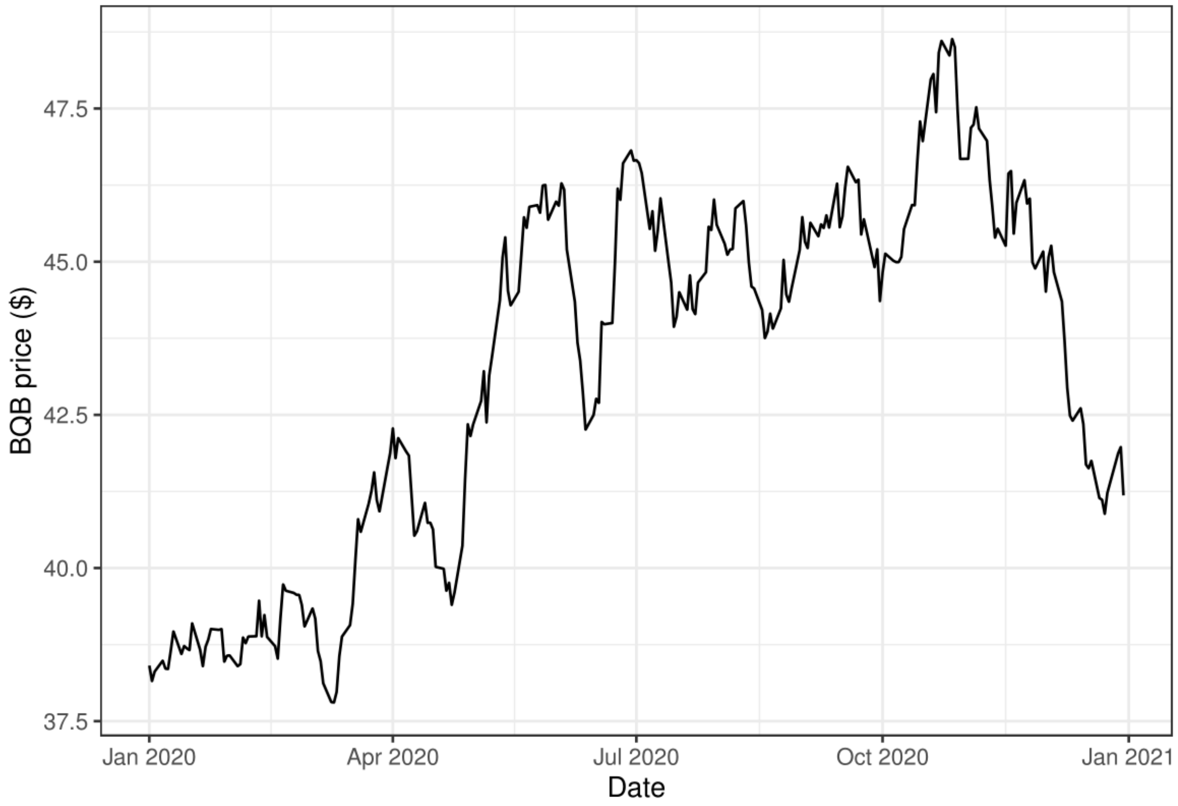
Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

```
bqb_stock_price_df%>%head(10)
```

```
##          date    price
## 1  2020-01-01 38.40823
## 2  2020-01-02 38.15537
## 3  2020-01-03 38.31118
## 4  2020-01-06 38.48808
## 5  2020-01-07 38.35830
## 6  2020-01-08 38.35286
## 7  2020-01-09 38.64673
## 8  2020-01-10 38.96761
## 9  2020-01-13 38.59588
## 10 2020-01-14 38.72828
```

Modelling a time series of stock prices

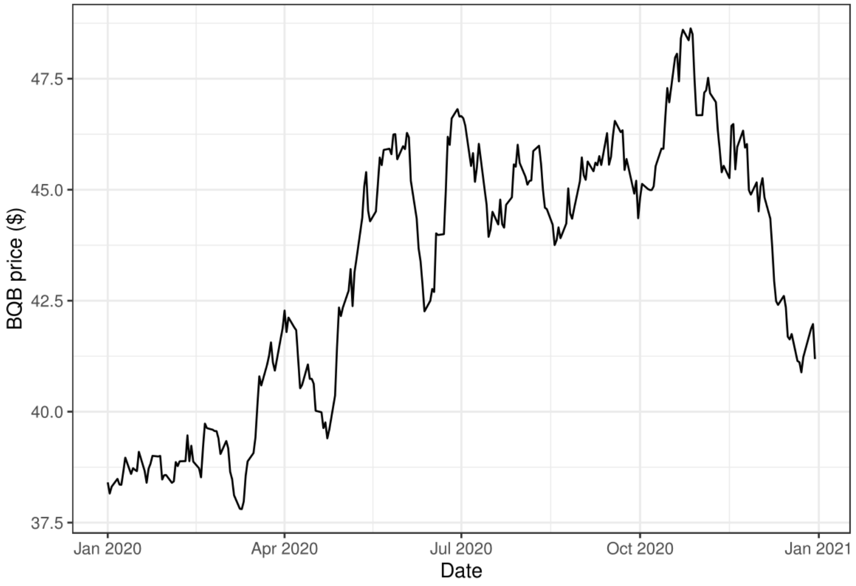
```
bqb_stock_price_df%>%
  ggplot(aes(x=date,y=price))+
  geom_line()+theme_bw()+
  ylab("BQB price ($)")+xlab("Date")
```



Modelling a time series of stock prices

Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.



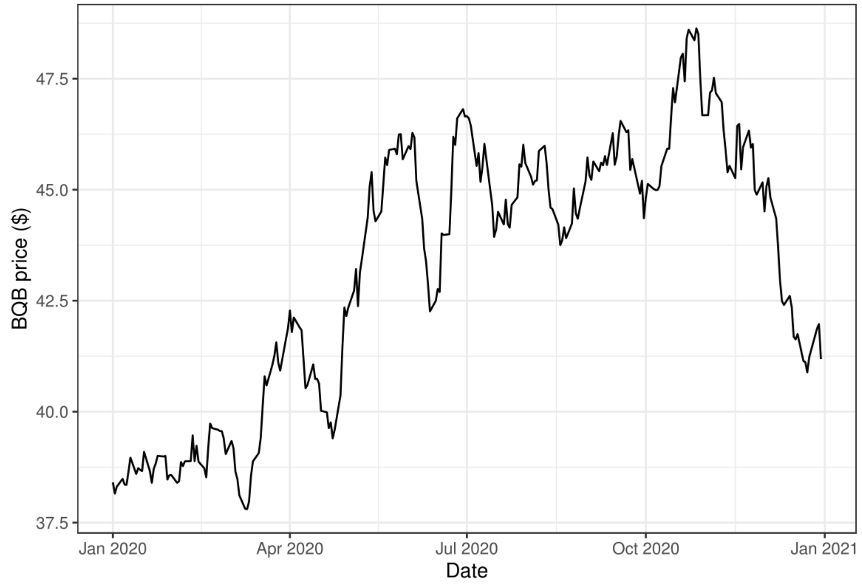
Modelling a time series of stock prices

Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

To see this let's look at the sample correlation between S_t and S_{t-1}.

```
bqb_stock_price_df%>%
  mutate(price_yesterday=lag(price))%>%
  select(price,price_yesterday)%>%
  cor(use="pairwise.complete.obs")
```



Modelling a time series of stock prices

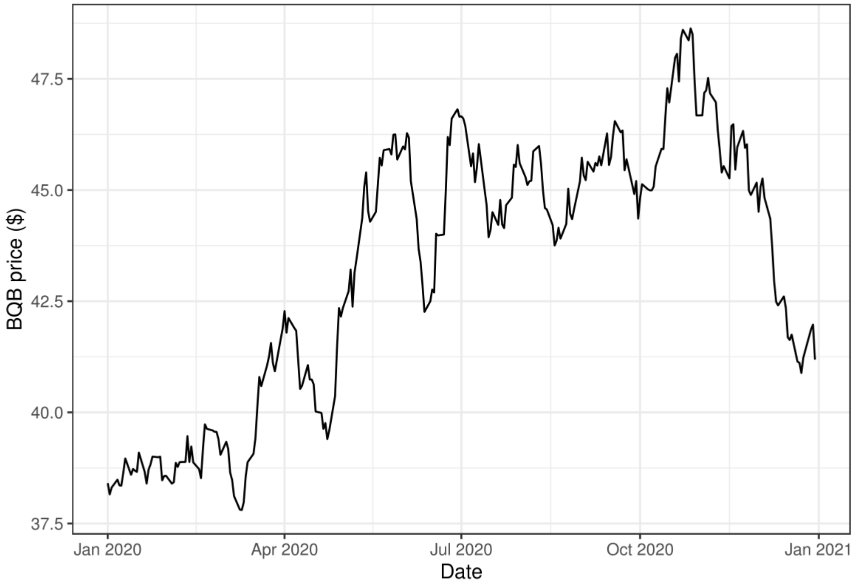
Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

To see this let's look at the sample correlation between S_t and S_{t-1}.

```
bqb_stock_price_df%>%
  mutate(price_yesterday=lag(price))%>%
  select(price,price_yesterday)%>%
  cor(use="pairwise.complete.obs")
```

```
##                     price price_yesterday
## price           1.0000000       0.9880581
## price_yesterday 0.9880581       1.0000000
```

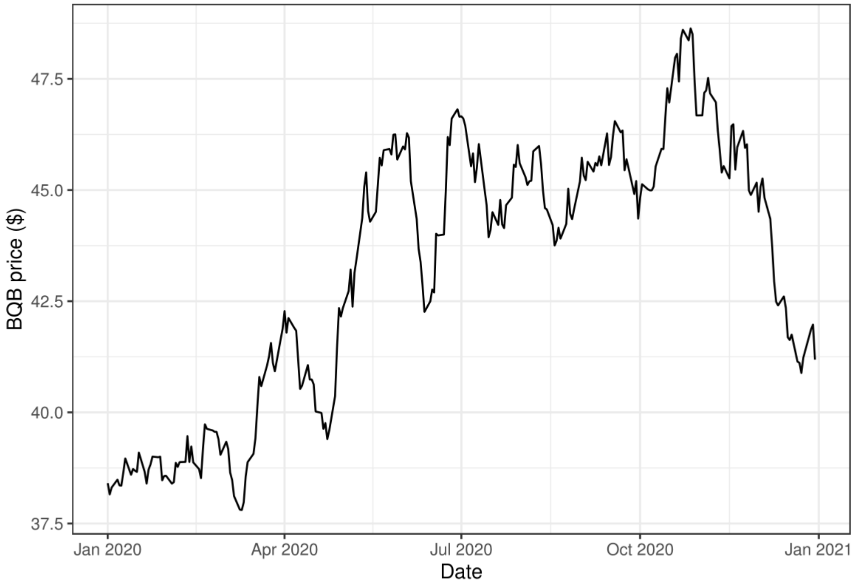


Modelling a time series of stock prices

Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

A simple model for stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_t \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.



Modelling a time series of stock prices

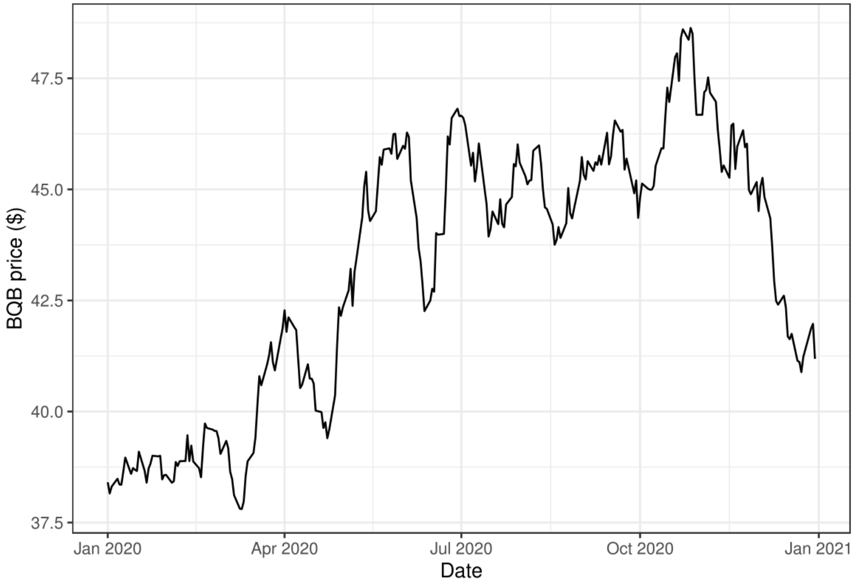
Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

A simple model for stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_t \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

The parameter μ corresponds to the degree of drift in the process.

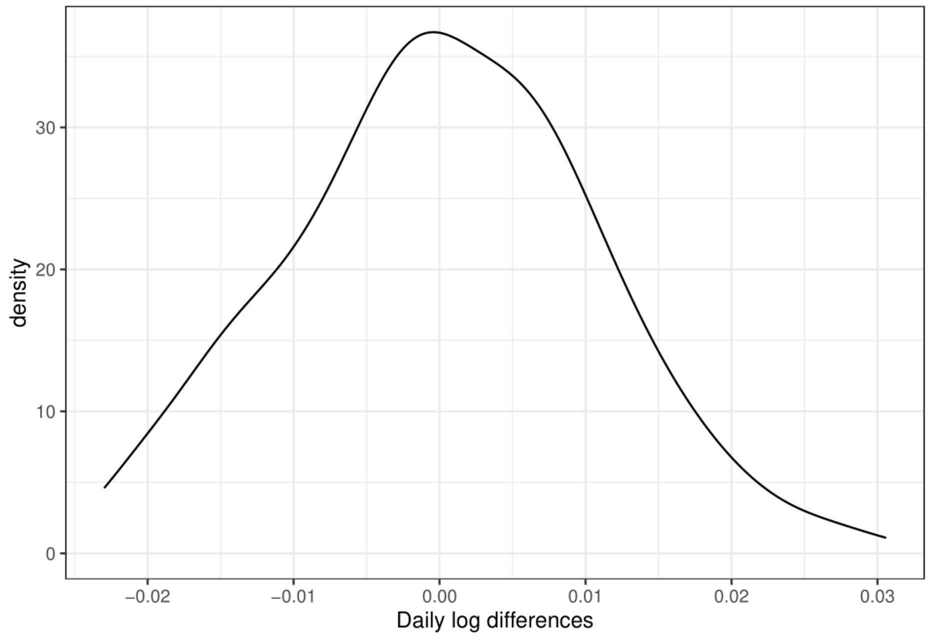
The parameter σ corresponds to the level of volatility.



Modelling a time series of stock prices

A simple model for stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_t \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

```
bqb_stock_price_df%>%
  mutate(log_diffs=log(price)-log(lag(price)))%>%
  ggplot(aes(x=log_diffs))+
  geom_density()+theme_bw()+
  xlab("Daily log differences")
```

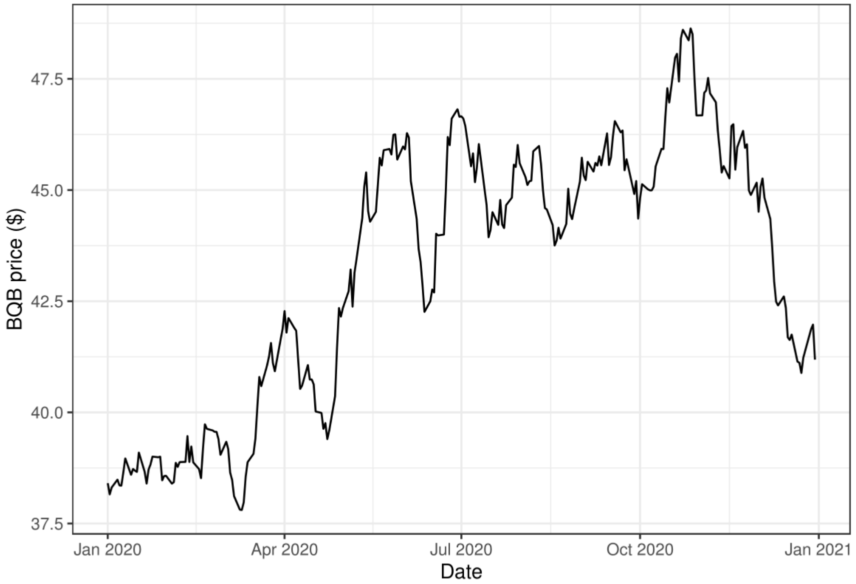


Modelling a time series of stock prices

Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

A simple model for stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_t \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.



Modelling a time series of stock prices

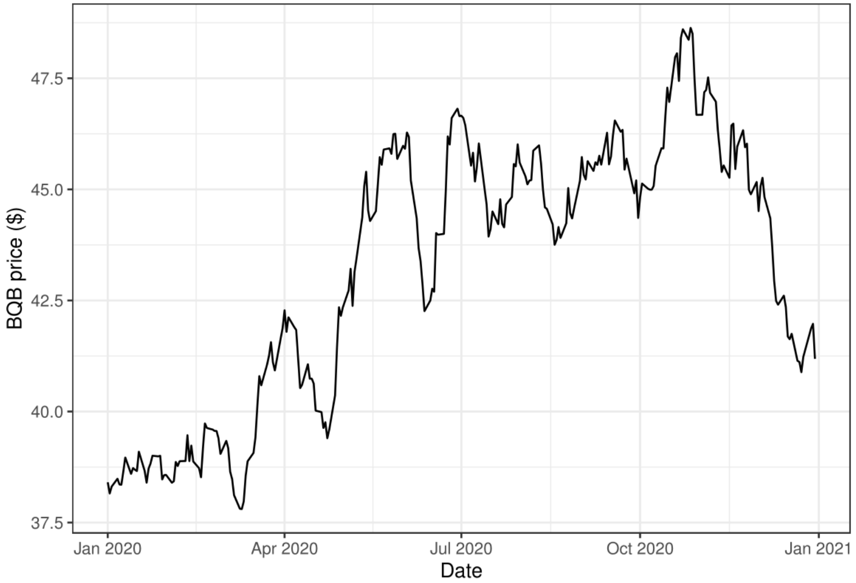
Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

A simple model for stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_t \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

The parameter μ corresponds to the degree of drift in the process.

The parameter σ corresponds to the level of volatility.



Modelling a time series of stock prices

Let's consider a time series of stock prices S_t for $t = 1, \ldots, 365$.

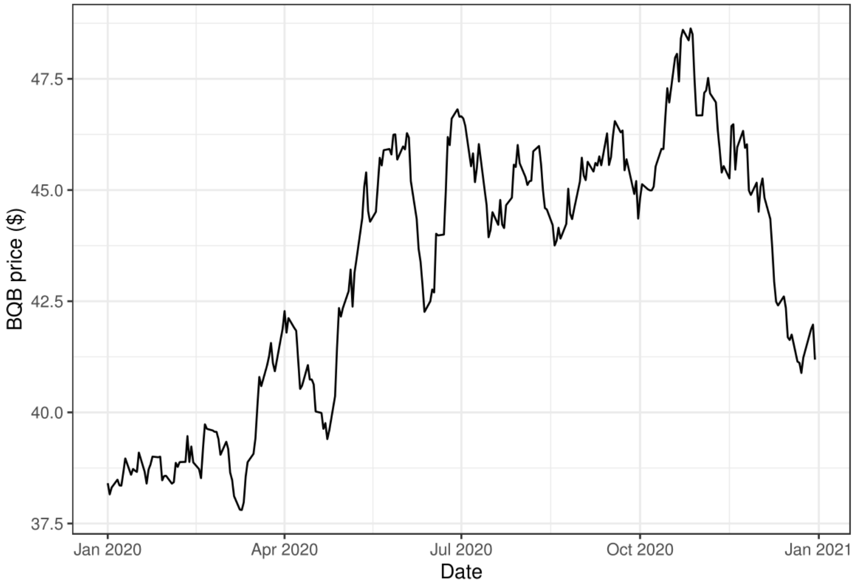
Notice that the series of prices $S_1, \ldots, S_{365}$ is not independent.

A simple model for stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_t \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

The parameter μ corresponds to the degree of drift in the process.

The parameter σ corresponds to the level of volatility.

How can we test hypotheses about the volatility parameter σ?



A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$;

... and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$;

... and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

The sample variance $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$ is a minimum variance unbiased estimator for σ^2.

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$;

... and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

The sample variance $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$ is a minimum variance unbiased estimator for σ^2.

A natural test statistic is

$$\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}.$$

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

A natural test statistic is $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}$ where $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$.

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

A natural test statistic is $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}$ where $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$.

If H_0 holds then $\mathbb{E}[S_n^2] = \sigma_0^2$ so $\mathbb{E}[\hat{\chi}^2] = (n-1) \cdot \mathbb{E}[S_n^2] \cdot (\sigma_0^2)^{-1} = n - 1$.

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

A natural test statistic is $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}$ where $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$.

If H_0 holds then $\mathbb{E}[S_n^2] = \sigma_0^2$ so $\mathbb{E}[\hat{\chi}^2] = (n-1) \cdot \mathbb{E}[S_n^2] \cdot (\sigma_0^2)^{-1} = n - 1$.

If $\hat{\chi}^2$ is much larger or much smaller than $n-1$, perhaps this is good evidence against the null hypothesis H_0.

A one sample test of population variance

Suppose we have an i.i.d. Gaussian sample $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to test the value of the population variance σ^2.

We have the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ and alternative hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$.

A natural test statistic is $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}$ where $S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$.

If H_0 holds then $\mathbb{E}[S_n^2] = \sigma_0^2$ so $\mathbb{E}[\hat{\chi}^2] = (n-1) \cdot \mathbb{E}[S_n^2] \cdot (\sigma_0^2)^{-1} = n - 1$.

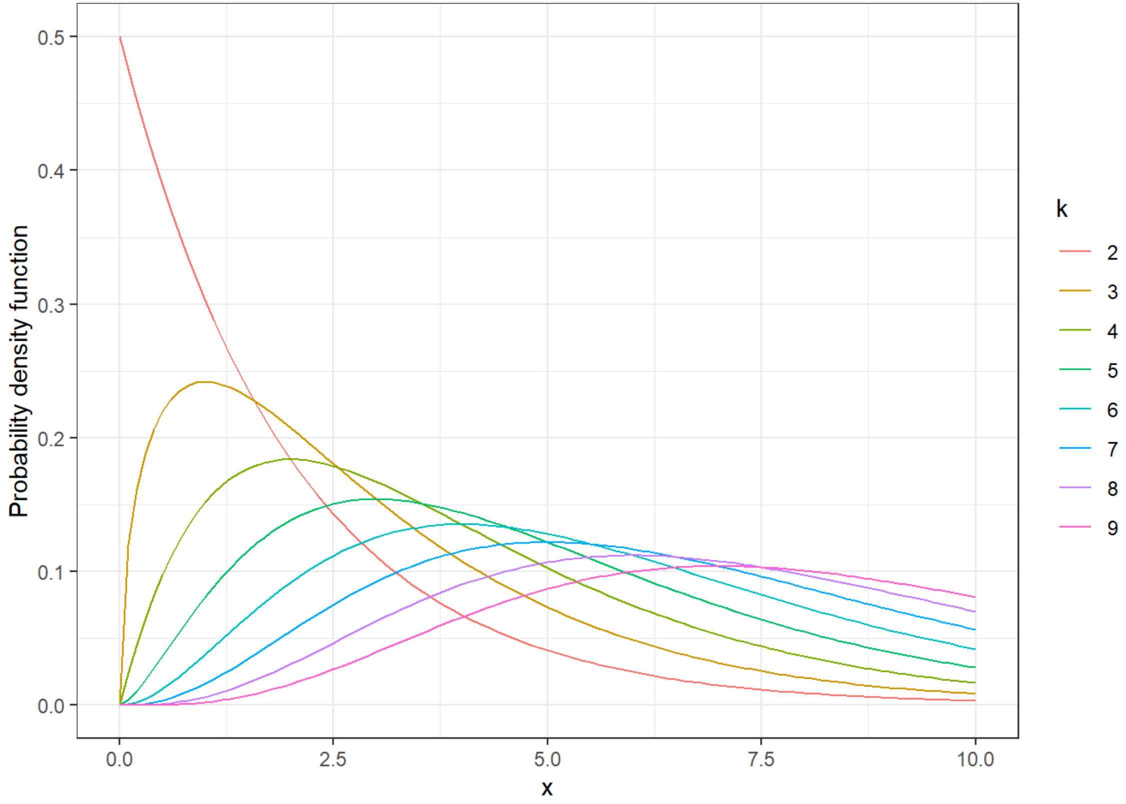
If $\hat{\chi}^2$ is much larger or much smaller than $n-1$, perhaps this is good evidence against the null hypothesis H_0.

Lemma (Cochran, 1934). *Suppose we have an i.i.d. Gaussian sample $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$. Then the chi-squared statistic $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$ is follows a chi-squared distribution with $n-1$ degrees of freedom.*

A one sample test of population variance

A random variable Q is said to be chi-squared with k degrees of freedom $Q \sim \chi^2(k)$ if

$Q = \sum_{i=1}^{k} Z_i^2$ with $Z_1, \cdots, Z_k \sim \mathcal{N}(0,1)$ independent and identically distributed.



A one sample test of population variance

Lemma (Cochran, 1934). *Suppose we have an i.i.d. Gaussian sample* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$. *Then the chi-squared statistic* $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$ *is follows a chi-squared distribution with* $n-1$ *degrees of freedom.*

Suppose that we observe a numerical value of x chi-squared statistic $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$.

A one sample test of population variance

Lemma (Cochran, 1934). *Suppose we have an i.i.d. Gaussian sample* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$. *Then the chi-squared statistic* $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$ *is follows a chi-squared distribution with* $n-1$ *degrees of freedom.*

Suppose that we observe a numerical value of x chi-squared statistic $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$.

The p-value is the probability of a obtaining a quantity at least as extreme as the observed value under H_0.

A one sample test of population variance

Lemma (Cochran, 1934). *Suppose we have an i.i.d. Gaussian sample* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$. *Then the chi-squared statistic* $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$ *is follows a chi-squared distribution with* $n-1$ *degrees of freedom.*

Suppose that we observe a numerical value of x chi-squared statistic $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$.

The p-value is the probability of a obtaining a quantity at least as extreme as the observed value under H_0.

Let $F_{\chi^2_{n-1}}$ be the cumulative distribution function of a χ^2 random variable with $n-1$ degrees of freedom.

A one sample test of population variance

Lemma (Cochran, 1934). *Suppose we have an i.i.d. Gaussian sample* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$. *Then the chi-squared statistic* $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$ *is follows a chi-squared distribution with* $n-1$ *degrees of freedom.*

Suppose that we observe a numerical value of x chi-squared statistic $\hat{\chi}^2 := \frac{(n-1)S_n^2}{\sigma_0^2}$.

The p-value is the probability of a obtaining a quantity at least as extreme as the observed value under H_0.

Let $F_{\chi^2_{n-1}}$ be the cumulative distribution function of a χ^2 random variable with $n-1$ degrees of freedom.

We compute the p-value by $p = 2 \cdot \min\left\{\mathbb{P}(\hat{\chi}^2 \leq x | H_0), \mathbb{P}(\hat{\chi}^2 \geq x | H_0)\right\} = 2 \min\left\{F_{\chi^2_{n-1}}(x), 1 - F_{\chi^2_{n-1}}(x)\right\}$.

A one sample test of population variance

We compute the p-value by $p = 2 \cdot \min\left\{\mathbb{P}(\hat{\chi}^2 \leq x | H_0), \mathbb{P}(\hat{\chi}^2 \geq x | H_0)\right\} = 2 \min\left\{F_{\chi^2_{n-1}}(x), 1 - F_{\chi^2_{n-1}}(x)\right\}$.

```
chi_square_test_one_sample_var<-function(sample,sigma_square_null){

  sample<-sample[!is.na(sample)]
  # remove any missing values
  n<-length(sample)
  # sample length
  chi_squared_statistic<-(n-1)*var(sample)/sigma_square_null
  # compute test statistic
  p_value<-2*min(pchisq(chi_squared_statistic,df=n-1),
                 1-pchisq(chi_squared_statistic,df=n-1))
  # compute the p-value

  return(p_value)
}
```

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$, where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

We want to test if the volatility parameter $\sigma = 1/100$.

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

We want to test if the volatility parameter $\sigma = 1/100$.

Our null hypothesis is $H_0 : \sigma = 1/100$ and our alternative is $H_1 : \sigma \neq 1/100$.

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

We want to test if the volatility parameter $\sigma = 1/100$.

Our null hypothesis is $H_0 : \sigma = 1/100$ and our alternative is $H_1 : \sigma \neq 1/100$.

We choose a significance level of $\alpha = 0.05$.

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

We want to test if the volatility parameter $\sigma = 1/100$.

Our null hypothesis is $H_0 : \sigma = 1/100$ and our alternative is $H_1 : \sigma \neq 1/100$.

We choose a significance level of $\alpha = 0.05$.

```
bqb_stock_prices%>%
  mutate(log_diffs=log(price)-log(lag(price)))%>%
  pull(log_diffs)%>%
  chi_square_test_one_sample_var(sample=.,sigma_square_null = (1/100)^2)
```

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

We want to test if the volatility parameter $\sigma = 1/100$.

Our null hypothesis is $H_0 : \sigma = 1/100$ and our alternative is $H_1 : \sigma \neq 1/100$.

We choose a significance level of $\alpha = 0.05$.

```
bqb_stock_prices%>%
  mutate(log_diffs=log(price)-log(lag(price)))%>%
  pull(log_diffs)%>%
  chi_square_test_one_sample_var(sample=.,sigma_square_null = (1/100)^2)
```

```
## [1] 0.2502084
```

Testing the volatility parameter

Suppose we model stock prices is given by $S_t = S_{t-1} \cdot \exp(X_t)$,
where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables. Hence, $X_t = \log(S_t) - \log(S_{t-1})$.

We want to test if the volatility parameter $\sigma = 1/100$.

Our null hypothesis is $H_0 : \sigma = 1/100$ and our alternative is $H_1 : \sigma \neq 1/100$.

We choose a significance level of $\alpha = 0.05$.

```
bqb_stock_prices%>%
  mutate(log_diffs=log(price)-log(lag(price)))%>%
  pull(log_diffs)%>%
  chi_square_test_one_sample_var(sample=.,sigma_square_null = (1/100)^2)
```

```
## [1] 0.2502084
```

The p-value exceeds the significance level so we cannot reject the null hypothesis.

What have we covered?

- We began with an illustrative time series example involving a stock price;
- We modelled the log differences could as a sequence of i.i.d. Gaussian random variables;
- We then considered testing the value of the population variance;
- We saw that the chi-squared statistic involving the sample variance follows a chi-squared distribution;
- We used this distributional behavior to derive the chi-squared test for the variance.
- In a future lectures we will consider a large family of hypothesis tests based on the chi-squared distribution.

University of BRISTOL

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Include EMATM0061 in the subject of your email.

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.