Hypothesis testing
for the population variance

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

University of BRISTOL

What we will cover today

We will consider the problem of hypothesis testing for the population variance

We will look at the use of chi-squared distributions for hypothesis testing

We will introduce the chi-squared test for population variance based on the distributional behaviour of sample variance

We will look at an illustrative time series example where our focus is the variance parameter

Today's focus

In our previous lectures, we discuss hypothesis testing for the population mean

- e.g., one sample t-test, paired t-test, Welch's t-test, $\cdots$

In this lecture, our interest is in the population variance...

Given a sample that is randomly drawn from a population, we want to know about the population variance. We want to decide a statement about the population variance is true or not with a hypothesis test.

A one-sample test of population variance

A one sample test of population variance

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

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

The hypotheses:

Null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given).

Alternative hypothesis: $H_1 : \sigma^2 \neq \sigma_0^2$.

The key question here is: what could be a suitable test statistic for this hypothesis-testing problem?

Recall that the test statistic is some function of the sample which:

i). has a known distribution under the null hypothesis H_0.

ii). often takes on large or "extreme" values under the alternative hypothesis H_1.

Test statistics

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

Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

Recall that the test statistic is some function of the sample which:

i). has a known distribution under the null hypothesis H_0.

ii). often takes on large or "extreme" values under the alternative hypothesis H_1.

Intuition: we may start from the sample variance:

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

We can define a test statistic as

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

If $\sigma \neq \sigma_0$, then $\hat{\chi}^2$ tends to be away from $n-1$, as ii) requires.

Chi-squared test statistics

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

Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

Recall that the test statistic is some function of the sample which:

i). has a known distribution under the null hypothesis H_0.

ii). often takes on large or "extreme" values under the alternative hypothesis H_1.

We can define a test statistic as $\hat{\chi}^2 := (n-1)\frac{S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2}$.

Do we know the distribution of $\hat{\chi}^2$ which is required by i)?

Yes! If H_0 is true, then $\hat{\chi}^2$ follows a chi-squared distribution (see the next slide).

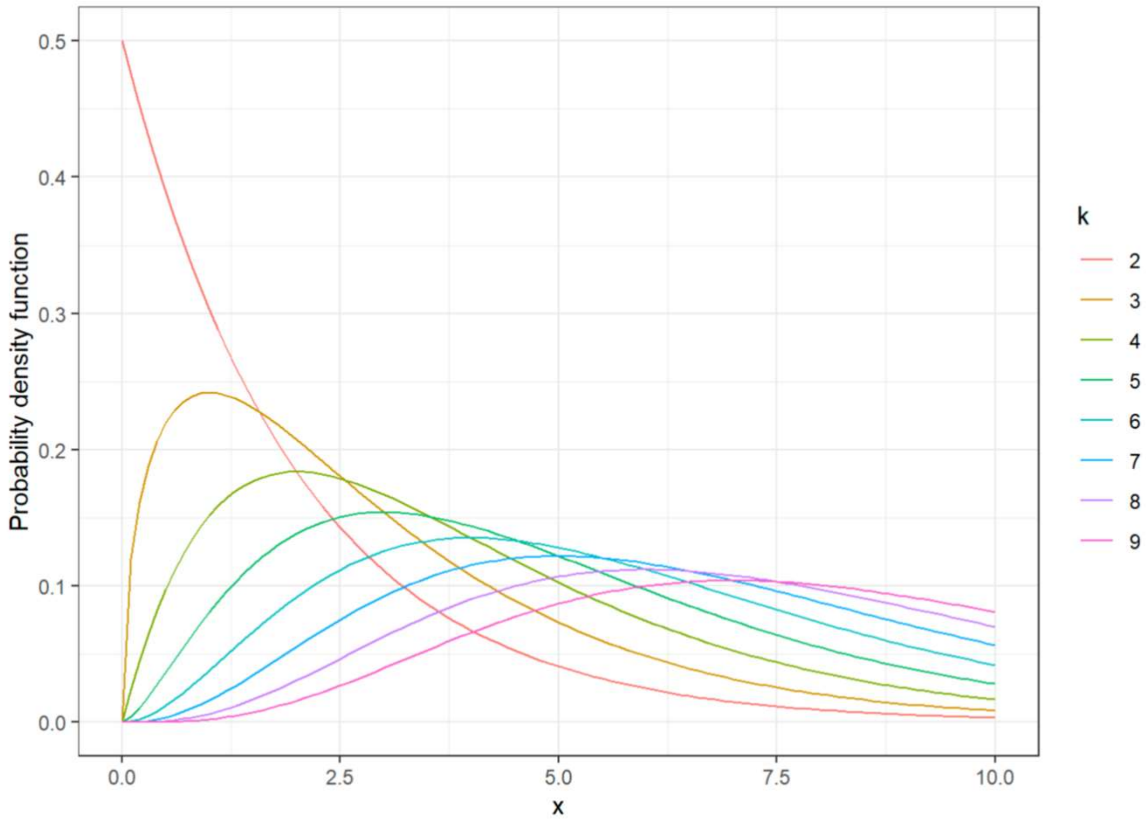
Lemma (Cochran, 1934)

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

Chi-squared distribution

A random variable Q is said to be chi-squared with k degrees of freedom if $Q = \sum_{i=1}^{k} Z_i^2$ with independent $Z_1, Z_2, \cdots, Z_k \sim \mathcal{N}(0,1)$.

We write $Q \sim \chi^2(k)$.



Expectation $\mathbb{E}(Q) = \sum_{i=1}^{k} \mathbb{E}\big(Z_i^2\big) = k.$

Variance test with chi-squared distribution: main idea

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

Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

We can define a test statistic as $\hat{\chi}^2 := (n-1)\frac{S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2}$.

If H_0 is true, then $\hat{\chi}^2$ is chi-squared distributed with $n-1$ degrees of freedom.

Then we can compute the numerical value of $\hat{\chi}^2$ based on our sample...

... and then compute the p-value with the numerical value

... and then draw the conclusion on the hypothesis test

Example

Modelling a time series of stock prices

Let's consider a time series of stock prices $S_1, S_2, \cdots, S_{365}$.

Suppose that we have the following sample stored in a data frame:

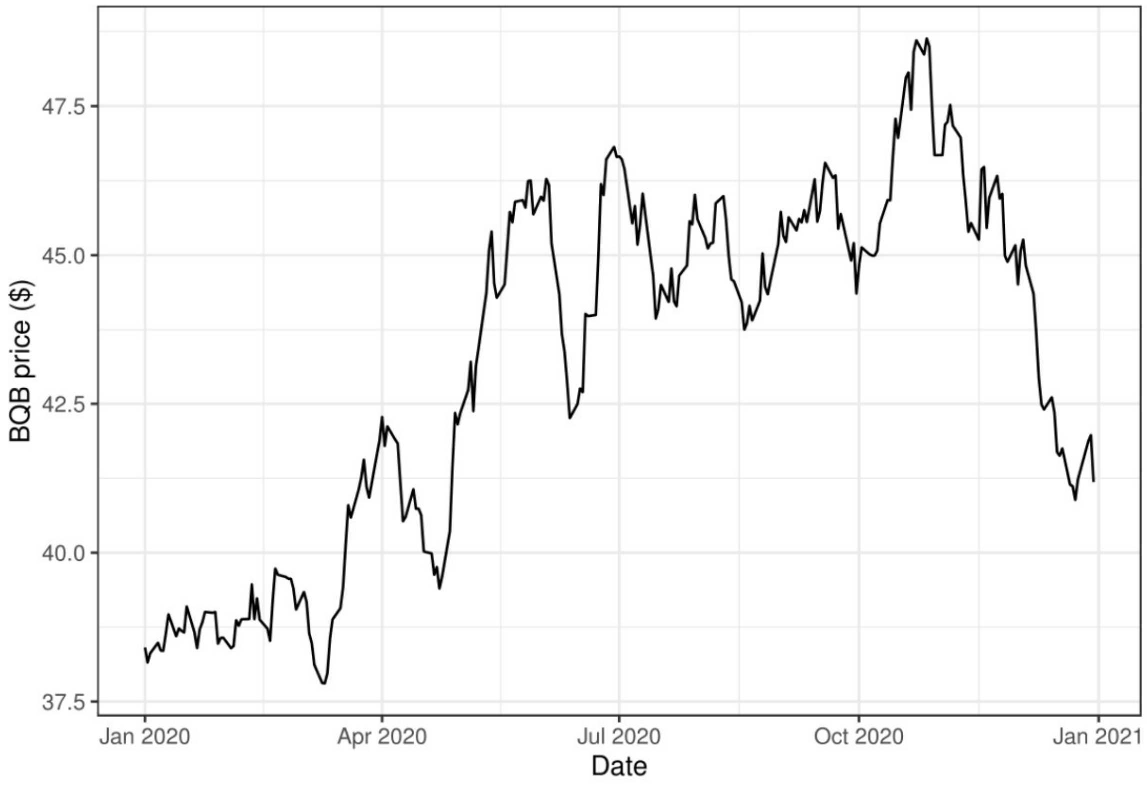
```
bqb_stock_price_df%>%head(10)
```

```
##          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

First, let's visualise the time series:

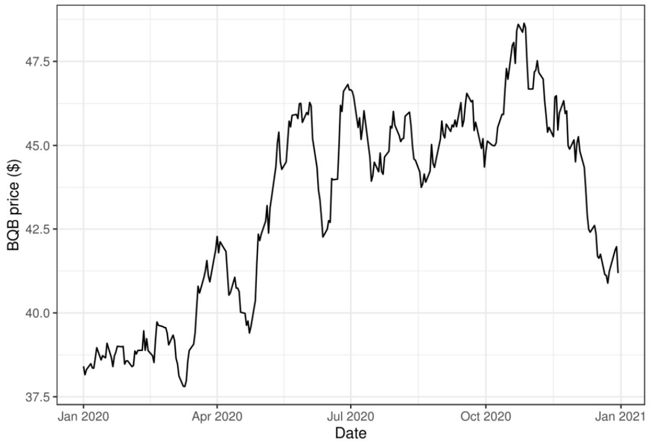
```
bqb_stock_price_df%>%
  ggplot(aes(x=date,y=price))+
  geom_line()+theme_bw()+
  ylab("BQB price ($)")+xlab("Date")
```



Modelling a time series of stock prices

Let's consider a time series of stock prices $S_1, S_2, \cdots, S_{365}$.

Notice that the series of price $S_1, \cdots, S_{365}$ is not independent, as the stock price today depends on the price yesterday (see the plot).



To see this, we can also 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
```

So S_t and S_{t-1} are correlated, hence they can not be independent.

Modelling a time series of stock prices

Let's consider a time series of stock prices $S_1, S_2, \cdots, S_{365}$.

Notice that the series of price $S_1, \cdots, S_{365}$ is not independent, as the stock price today depends on the price yesterday.

Given the dependency, we can model the stock price by

$$S_t := S_{t-1} \cdot \exp(X_t)$$

where $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

Here, we are interested in the change of prices, so we investigate the random variables X_t.

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

– The parameter σ corresponds to the level of volatility.

Question: How can we test hypotheses about the volatility parameter σ?

Statistical hypothesis testing: key stages

Suppose we have a clear research hypothesis and some high-quality data from a well-designed experiment.

The key stages of statistical hypothesis testing are as follows:

1. Form our statistical hypothesis including a null hypothesis and an alternative hypothesis.
2. Apply model checking to validate any modelling assumptions.
3. Choose our desired significance level.
4. Select an appropriate statistical test.
5. Compute the numerical value of the test statistic from the data.
6. Compute a p-value based on the test statistic.
7. Draw conclusions based on the relationship between the p-value and the significance level.

Formulating the hypothesis test

We can model the stock price by $S_t := S_{t-1} \cdot \exp(X_t)$ where $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

1. Form our statistical hypothesis including a null hypothesis and an alternative hypothesis.

Now, our sample is given by $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

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

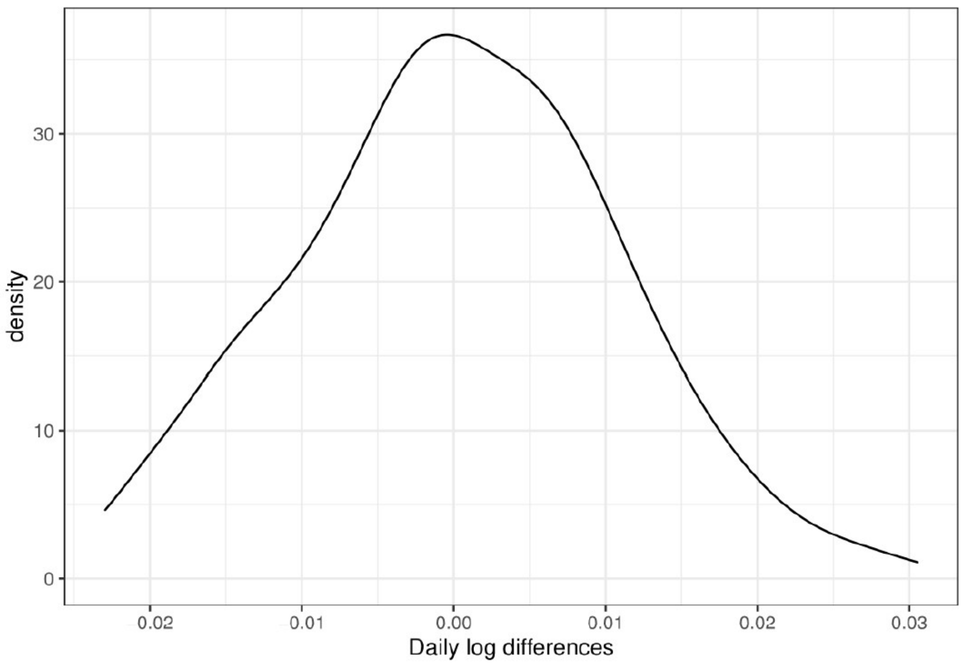
Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

Checking modelling assumption

We can model the stock price by $S_t := S_{t-1} \cdot \exp(X_t)$ where $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Gaussian random variables.

2. Apply model checking to validate any modelling assumptions.

```
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")
```



$$\log(S_t) = \log(S_{t-1}\exp(X_t)) \\ = \log(S_{t-1}) + X_t$$

$$X_t = \log(S_t) - \log(S_{t-1})$$

Significance level and test statistic

Now, our sample is given by $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

3. Choose our desired significance level.

Next we choose a significance level: $\alpha = 0.05$

4. Select an appropriate statistical test.

This is a one-sample test of population variance with Gaussian data assumption!

Therefore, we can use the test statistic $\hat{\chi}^2 := (n-1)\frac{S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2}$, as discussed previously.

p-value

Now, our sample is given by $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

Therefore, we can use the test statistic $\hat{\chi}^2 := (n-1)\frac{S_n^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}$, as discussed previously.

5. Compute the numerical value of the test statistic from the data.

Suppose that the numerical value of the test statistic is x.

6. Compute a p-value based on the test statistic.

The p-value is the probability of 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 χ^2 random variable with $n-1$ degrees of freedom.

We compute the p-value by

$$p := 2 \cdot \min\left(\mathbb{P}(\hat{\chi}^2 \leq x \mid H_0), \mathbb{P}(\hat{\chi}^2 \geq x \mid H_0)\right) = 2\min\left(F_{\chi^2_{n-1}}(x), 1 - F_{\chi^2_{n-1}(x)}\right).$$

p-value

We compute the p-value by

$$p := 2 \cdot \min\left(\mathbb{P}(\hat{\chi}^2 \leq x \mid H_0), \mathbb{P}(\hat{\chi}^2 \geq x \mid H_0)\right) = 2\min\left(F_{\chi^2_{n-1}}, 1 - F_{\chi^2_{n-1}}\right).$$

```
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)
}
```

5. Compute the numerical value of the test statistic from the data.
6. Compute a p-value based on the test statistic.

Testing the volatility parameter

Null $H_0 : \sigma^2 = \sigma_0^2$ (here σ_0^2 is given). Alternative: $H_1 : \sigma^2 \neq \sigma_0^2$.

Now, we carry out a population test below. Here we take $\sigma_0 = 1/100$.

```
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
```

7. Draw conclusions based on the relationship between the p-value and the significance level.

Conclusion: The p-value is bigger than the significance level, so we can not reject the null hypothesis.

What have we covered?

We considered the problem of one-sample hypothesis test for population variance

We derived a test statistic from the sample variance.

The test statistic follows a chi-square distribution

We investigated a time series example involving a stock price.

We study the volatility parameter with a population variance test.

Thanks for listening!

Dr. Rihuan Ke
rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science