Probability and Statistics course

Introduction

Gabor Vigh

TTK Department of Probability Theory and Statistics
ELTE

November 30, 2025

Week 12

Source: Slides created based on [Baron, 2014]

Inference about variances

We'll derive confidence intervals and tests for the population variance:

a variance is a scale and not a location parameter,

b the distribution of its estimator, the sample variance, is not symmetric.

Variance often needs to be estimated or tested for quality control, in order to assess stability and accuracy, evaluate various risks, and also, for tests and confidence intervals for the population means when variance is unknown. The sample variance formula is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The summands $(X_i - \bar{X})^2$ are not quite independent, as the Central Limit Theorem requires, because they all depend on $\bar{X}$. Nevertheless, the distribution of s^2 is approximately Normal, under mild conditions, when the sample is large. For small to moderate samples, the distribution of s^2 is not Normal at all. It is not even symmetric. Indeed, why should it be symmetric if s^2 is always non-negative!

Distribution of the sample variance

Chi-square distribution

When observations $X_1, \ldots, X_n$ are independent and Normal with $\text{Var}(X_i) = \sigma^2$, the distribution of

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

is Chi-square with $(n-1)$ degrees of freedom.

The probability density function (PDF) for a Chi-square random variable X with ν degrees of freedom is:

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \quad x > 0$$

where Γ is the Gamma function.

The expected value (mean) and variance are:

$$E(X) = \nu$$

$$\text{Var}(X) = 2\nu$$

Confidence interval for the population variance

Let us construct a $(1-\alpha)100\%$ confidence interval for the population variance σ^2, based on a sample of size n. As always, we start with the estimator, the sample variance s^2.

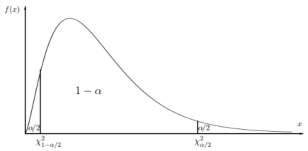


Figure: Critical values of the Chi-square distribution

A rescaled sample variance $\frac{(n-1)s^2}{\sigma^2}$ has χ^2 density

$$P\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) = 1-\alpha.$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha.$$

Confidence interval for the population variance and std

Confidence interval for the variance

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

Confidence interval for the standard deviation

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right]$$

Example

There is a sample containing n = 6 measurements, 2.5, 7.4, 8.0, 4.5, 7.4, and 9.2. Give confidence interval for the std deviation!

Testing variance

Table: Summary of Chi-square Test for Population Variance

Null H_0	**Alternative** H_A	**Test Statistic**	**Rejection Region**	***P*-value Computation**
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2_{\text{obs}} = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2_{\text{obs}} \geq \chi^2_{\alpha}$	$P\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}$
	$\sigma^2 < \sigma_0^2$		$\chi^2_{\text{obs}} \leq \chi^2_{1-\alpha}$	$P\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\}$
	$\sigma^2 \neq \sigma_0^2$		$\chi^2_{\text{obs}} \geq \chi^2_{\alpha/2}$ or $\chi^2_{\text{obs}} \leq \chi^2_{1-\alpha/2}$	$2\min\left(P\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}, P\left\{\chi^2 \leq \chi^2_{\text{obs}}\right\}\right)$

Example

There is a sample containing n = 6 measurements, 2.5, 7.4, 8.0, 4.5, 7.4, and 9.2. Calculate P-value whether the std dev is 2.2!

Comparison of two variances. F-distribution

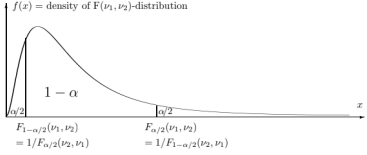
To compare variances or standard deviations, two independent samples $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ are collected, one from each population. Unlike population means or proportions, variances are scale factors, and they are compared through their ratio $\theta = \frac{\sigma_X^2}{\sigma_Y^2}$. A natural estimator for the ratio of population variances $\theta = \sigma_X^2/\sigma_Y^2$ is the ratio of sample variances $\hat{\theta} = \frac{s_X^2}{s_Y^2} = \frac{\sum(X_i - \bar{X})^2/(n-1)}{\sum(Y_i - \bar{Y})^2/(m-1)}$. The distribution of this statistic is simply called **F-distribution** with $(n-1)$ and $(m-1)$ degrees of freedom.

Distribution of the ratio of sample variances

For independent samples $X_1, \ldots, X_n$ from Normal(μ_X, σ_X) and $Y_1, \ldots, Y_m$ from Normal(μ_Y, σ_Y), the standardized ratio of variances

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{\sum(X_i - \bar{X})^2/\sigma_X^2/(n-1)}{\sum(Y_i - \bar{Y})^2/\sigma_Y^2/(m-1)}$$

has F-distribution with $(n-1)$ and $(m-1)$ degrees of freedom. If F has $F(\nu_1, \nu_2)$ distribution, then the distribution of $\frac{1}{F}$ is $F(\nu_2, \nu_1)$.



Confidence interval for the ratio of population variances

Start with the estimator, $\hat{\theta} = s_X^2/s_Y^2$. Standardizing it to

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{s_X^2/s_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{\hat{\theta}}{\theta},$$

we get an F-variable with $(n-1)$ and $(m-1)$ degrees of freedom. Therefore,

$$P\left(F_{1-\alpha/2}(n-1, m-1) \leq \frac{\hat{\theta}}{\theta} \leq F_{\alpha/2}(n-1, m-1)\right) = 1-\alpha,$$

Solving the double inequality for the unknown parameter θ, we get

$$P\left(\frac{\hat{\theta}}{F_{\alpha/2}(n-1, m-1)} \leq \theta \leq \frac{\hat{\theta}}{F_{1-\alpha/2}(n-1, m-1)}\right) = 1-\alpha.$$

Therefore, the $(1-\alpha)100\%$ confidence interval for the ratio of variances $\theta = \frac{\sigma_X^2}{\sigma_Y^2}$ is:

$$\left[\frac{\hat{\theta}}{F_{\alpha/2}(n-1, m-1)}, \frac{\hat{\theta}}{F_{1-\alpha/2}(n-1, m-1)}\right] = \left[\frac{s_X^2/s_Y^2}{F_{\alpha/2}(n-1, m-1)}, \frac{s_X^2/s_Y^2}{F_{1-\alpha/2}(n-1, m-1)}\right]$$

The critical values of $F(\nu_1, \nu_2)$ and $F(\nu_2, \nu_1)$ distributions are related as follows:

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_{\alpha}(\nu_2, \nu_1)}$$

The confidence interval is

$$\left[\frac{s_X^2}{s_Y^2 F_{\alpha/2}(n-1, m-1)}, \quad \frac{s_X^2 F_{\alpha/2}(m-1, n-1)}{s_Y^2}\right]$$

F-tests comparing two variances

Table: Summary of F-Test for Comparing Two Population Variances

Null $\mathbf{H_0}: \frac{\sigma_X^2}{\sigma_Y^2} = \theta_0$	**Alternative** $\mathbf{H_A}$	**Rejection Region**	P-**value**
θ_0	$\frac{\sigma_X^2}{\sigma_Y^2} > \theta_0$	$F_{\text{obs}} \geq F_{\alpha}(n-1, m-1)$	$P\{F \geq F_{\text{obs}}\}$
	$\frac{\sigma_X^2}{\sigma_Y^2} < \theta_0$	$F_{\text{obs}} \leq F_{1-\alpha}(n-1, m-1)$	$P\{F \leq F_{\text{obs}}\}$
	$\frac{\sigma_X^2}{\sigma_Y^2} \neq \theta_0$	$F_{\text{obs}} \geq F_{\alpha/2}(n-1, m-1)$ or $F_{\text{obs}} \leq F_{1-\alpha/2}(n-1, m-1)$	$2\min\left(P\{F \geq F_{\text{obs}}\}, P\{F \leq F_{\text{obs}}\}\right)$

Test Statistic: $F_{\text{obs}} = \frac{s_X^2}{s_Y^2}/\theta_0$, where F has $F(n-1, m-1)$ distribution.

Example

A data channel has the average speed of 180 Megabytes per second. A hardware upgrade is supposed to improve stability of the data transfer while maintaining the same average speed. Stable data transfer rate implies low standard deviation.

- **Before Upgrade (Population 1, or X):** Sample size $n_1 = 27$ Sample standard deviation $s_1 = 22$ Mbps
- **After Upgrade (Population 2, or Y):** Sample size $n_2 = 16$ Sample standard deviation $s_2 = 14$ Mbps

We are asked to construct a **90% confidence interval** for the relative change in the standard deviation ($\frac{\sigma_1}{\sigma_2}$) (assume Normal distribution of the speed). Can we infer that the channel became twice as stable as it was, if the increase of stability is measured by the proportional reduction of standard deviation?

Examples

1. We want to investigate whether the daily mean temperature in Budapest on October 18th was **below** 15°**C**. The daily mean temperatures from the past 4 years were as follows:$14.8, 12.2, 16.8, 11.1\ ^\circ$C Assume that the data originates from a Normal distribution.
   a. Write down the **null and alternative hypotheses**.
   b. Assume the population standard deviation is known: $\sigma = 2$. Test the hypothesis using a **significance level of** $\alpha = 0.05$.
      - Specify the **critical region** and the p-**value**.
      - What is the decision?
   c. Test the hypothesis **without using the prior information about the standard deviation** ($\sigma = 2$).
   d. What hypotheses should be formulated if we want to investigate whether the daily mean temperature in Budapest on October 18th was **different from** 15°**C**? Test this hypothesis using the given data.

Given Critical Values: $(z_{0.05} = -1.645, \quad \Phi(1.275) \approx 0.899, \quad t_{3;0.05} = -2.353, \quad z_{0.975} = 1.96)$

Example

2 The two samples below relate to the defect rates (in per mille) observed in two different factory units. Can we state that factory unit "A" performed better? (We can assume that the samples are normally distributed and independent.)

Unit A	11.9	12.1	12.8	12.2	12.5	11.9	12.5	11.8	12.4	12.9
Unit B	12.1	12.0	12.9	12.2	12.7	12.6	12.6	12.8	12.0	13.1

Given Critical Values:

$$(F_{9,9;0.975} \approx 4.026, \quad t_{18;0.05} \approx -1.734)$$

Example

3 Two servers were compared. The average running time for 30 executions on the first server was 6.7 seconds, while, independently, the average running time for 20 executions on the second server was 7.2 seconds. Investigate whether there is a significant difference between the speeds of the two servers, assuming the standard deviation of the running times was 0.5 seconds on both machines. ($z_{0.975} = 1.96$)

4 The two samples below contain concentration data for an air pollutant found in the atmosphere at 10 busy intersections. The first row contains the figures for November 15th, and the second row contains the figures for November 29th. Has the air pollution level significantly changed?

Date	**Concentration Data**									
November 15th (X)	20.9	17.1	15.8	18.8	20.1	15.6	14.8	24.1	18.9	12.5
November 29th (Y)	21.4	16.7	16.4	19.2	19.9	16.6	15.0	24.0	19.2	13.2

Given Critical Value:

$$(t_{9;0.975} = 2.262)$$

Chi-square tests - Testing a distribution

The **Chi-square statistic** (χ^2) is defined as:

$$\chi^2 = \sum_{k=1}^{N} \frac{\{\text{Obs}(k) - \text{Exp}(k)\}^2}{\text{Exp}(k)}$$

- N define # of categories or groups of data defined depending on our testing problem
- $\text{Obs}(k)$ is the actually observed number of sampling units in category k
- $\text{Exp}(k) = E\{\text{Obs}(k)|H_0\}$ is the expected number of sampling units in category k if the null hypothesis H_0 is true.

Useful thoughts:

- always a **one-sided, right-tail test** == large number when the 2 dist is close
- level α rejection region for this chi-square test is $R = [\chi^2_\alpha, +\infty)$
- the P-value is always calculated as $P = P\left\{\chi^2 \geq \chi^2_{\text{obs}}\right\}$.
- the **rule of thumb** requires an expected count of at least 5 in each category

Example

Example

Suppose that after losing a large amount of money, an unlucky gambler questions whether the game was fair and the die was really unbiased. The last 90 tosses of this die gave the following results:

Category	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Number of times it occurred (Obs(k))	20	15	12	17	9	17	90

We are asked to test whether the die was truly unbiased.

Testing independence

Apparently, chi-square statistics can help us test independence

$$H_0 : \text{Factors A and B are independent} \quad \text{vs} \quad H_A : \text{Factors A and B are dependent.}$$

Factors A and B are independent if any randomly selected unit x of the population belongs to categories A_i and B_j independently of each other.

$$H_0 : P\{x \in A_i \cap B_j\} = P\{x \in A_i\}P\{x \in B_j\} \quad \text{for all } i, j \text{ vs } H_A : P\{x \in A_i \cap B_j\} \neq P\{x \in A_i\}P\{x \in B_j\} \quad \text{for some } i, j.$$

To test these hypotheses, we collect a sample of size n and count n_{ij} units that landed in the intersection of categories A_i and B_j. These are the observed counts, which can be nicely arranged in a **contingency table**:

	B_1	B_2	$\cdots$	B_m	**row total**
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	$n_{\cdot\cdot} = n$

Notation $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$ is quite common for the row totals and column totals.

Testing independence

Then we estimate all the probabilities:

$$\hat{P}\{x \in A_i \cap B_j\} = \frac{n_{ij}}{n}, \quad \hat{P}\{x \in A_i\} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}, \quad \hat{P}\{x \in B_j\} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{\cdot j}}{n}.$$

If H_0 is true, then we can also estimate the probabilities of intersection as

$$\hat{P}_e\{x \in A_i \cap B_j\} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}$$

and estimate the expected counts as

$$\text{Exp}(i,j) = n\left(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right) = \frac{(n_{i\cdot})(n_{\cdot j})}{n} \; ; \; \text{Obs}(i,j) = \frac{n_{ij}}{n}$$

Chi-square test for independence

$$\text{Test statistic} \quad \chi^2_{\text{obs}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\{\text{Obs}(i,j) - \text{Exp}(i,j)\}^2}{\text{Exp}(i,j)},$$

where

- $\text{Obs}(i,j) = n_{ij}$ are observed counts,
- $\text{Exp}(i,j) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$ are estimated expected counts,
- and χ^2_{obs} has $(k-1)(m-1)$ d.f.

As always in this section, this test is one-sided and right-tail.

Example

Example

Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam, and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam **depends on whether it contains images or not**. The following data were collected on $n = 1000$ random email messages:

Observed Counts (Contingency Table)

$\text{Obs}(i,j) = n_{ij}$	**With images**	**No images**	$n_{i\cdot}$
Spam	160	240	400
No spam	140	460	600
$n_{\cdot j}$	300	700	1000

Least squares estimators

Definition

Response or dependent variable (Y) is a variable of interest that we predict based on one or several predictors.
Predictors or independent variables ($X^{(1)}, \ldots, X^{(k)}$) are used to predict the values and behavior of the response variable Y.
Regression of Y on $X^{(1)}, \ldots, X^{(k)}$ is the conditional expectation,

$$G(x^{(1)}, \ldots, x^{(k)}) = E\left\{ Y \mid X^{(1)} = x^{(1)}, \ldots, X^{(k)} = x^{(k)} \right\}.$$

It is a function of $x^{(1)}, \ldots, x^{(k)}$ whose form can be estimated from data.

Definition

Residuals

$$e_i = y_i - \hat{y}_i$$

are differences between observed responses y_i and their fitted values $\hat{y}_i = \hat{G}(x_i)$.
Method of least squares finds a regression function $\hat{G}(x)$ that minimizes the sum of squared residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Linear regression

The **Linear regression model** assumes that the conditional expectation

$$G(x) = E\{Y \mid X = x\} = \beta_0 + \beta_1 x$$

is a linear function of x. As any linear function, it has an intercept β_0 and a slope β_1. We minimize the sum of squared residuals

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{G}(x_i)\right)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We can do it by taking partial derivatives of Q, equating them to 0, and solving the resulting equations for β_0 and β_1. The partial derivatives are

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i);$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i.$$

Linear regression

The Ordinary Least Squares (OLS) estimators for the intercept ($\hat{\beta}_0$) and the slope ($\hat{\beta}_1$) are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \ , \ \hat{\beta}_1 = S_{xy}/S_{xx}$$

where S_{xx} and S_{xy} are the sums of squares: $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

The total variation among observed responses is measured by the **Total Sum of Squares**:

$$SS_{TOT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2.$$

A portion of this total variation is attributed to predictor X and the regression model connecting predictor and response. This portion is measured by the **Regression Sum of Squares**:

$$SS_{REG} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_1^2 S_{xx}.$$

The rest of total variation is attributed to "error." It is measured by the **Error Sum of Squares**:

$$SS_{ERR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

This is the portion of total variation not explained by the model. It equals the sum of squared residuals that the method of least squares minimizes. Thus, applying this method, we minimize the error sum of squares.
$SS_{TOT} = SS_{REG} + SS_{ERR}$

Goodness of Fit

The goodness of fit, appropriateness of the predictor and the chosen regression model can be judged by the proportion of SSTOT that the model can explain.

Definition

R-square, or **coefficient of determination** is the proportion of the total variation explained by the model, $R^2 = \frac{SS_{REG}}{SS_{TOT}}$.

It is always between 0 and 1, with high values generally suggesting a good fit.

Inference about the regression slope

Sampling distribution of a regression slope $\hat{\beta}_1$ is Normal$(\mu_{\hat{\beta}_1}, \sigma_{\hat{\beta}_1})$, where

$$\mu_{\hat{\beta}_1} = E(\hat{\beta}_1) = \beta_1$$

$$\sigma_{\hat{\beta}_1} = \text{Std}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$$

We estimate the standard error of $\hat{\beta}_1$ by

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}, \; s = \sqrt{\frac{SS_{ERR}}{n-2}}$$

and therefore, use T-intervals and T-tests.

Following the general principles, a $(1-\alpha)100\%$ confidence interval for the slope β_1 is

$$\text{Estimator} \pm t_{\alpha/2} \left(\begin{array}{c} \text{estimated} \\ \text{st. deviation} \\ \text{of the estimator} \end{array} \right) = \hat{\beta}_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}.$$

Testing hypotheses $H_0 : \beta_1 = B$ about the regression slope, use the T-statistic $t = \frac{\hat{\beta}_1 - B}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - B}{s/\sqrt{S_{xx}}}$. P-values, acceptance and rejection regions are computed based on T-distribution with $n2$ degrees of freedom.

Example

Example

A computer manager needs to know how the efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes, she gets the following results:

Data size (gigabytes), x	6	7	7	8	10	10	15
Processed requests, y	40	55	50	41	17	26	16

In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. The response variable here is the number of processed requests (y), and we attempt to predict it from the size of a data set (x).

References


Baron, M. (2014).
Probability and statistics for computer scientists.
CRC Press.