

ECON3150/4150 Spring 2015

Lecture 2 - Estimators and hypothesis testing

Siv-Elisabeth Skjelbred

University of Oslo

22. januar 2016

Last updated January 20, 2016

Overview

In this lecture we will cover remainder of chapter 2 and chapter 3

- Distributions
- Learn about estimators and their properties.
- Go through the statistical methods for estimation, hypothesis testing and confidence intervals.

Distributions

Probability distribution

Probability distribution

Probability distribution of a discrete random variable is a list of all possible pairs $[x, pr(x)]$ where x is the realizations of a random variable X and $pr(x)$ is the probability that x occurs.

- Important property: $\sum Pr(x) = 1$
- For a continuous variable you cannot list the probability of each event as there are too many and chance of any given outcome happening is 0.

Probability density functions - The normal distribution:

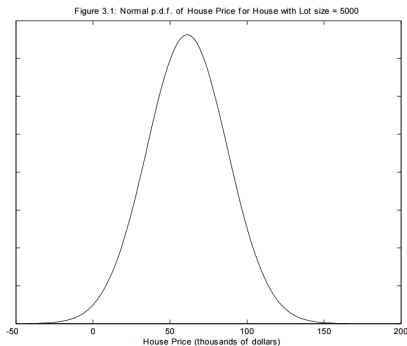
Normal distribution is one of the most common density functions.

Properties:

- Bell-shaped with (μ) and variance σ^2 , written as: $N(\mu, \sigma^2)$
- Symmetric around the mean
- 95% of its probability between $\mu + / - 1.96\sigma$.
- Note: A sum of n normally distributed random variables is itself normally distributed.

Example normal distribution

Consider again housing prices given that $X=5\ 000$.



- Bell shaped: The curve is highest for the most plausible values the house price might take.
- The mean (or average price of a house with a lot size of 5000 sqft is \$61 153.
- The variance is 683 812 (not much intuitive interpretation but it reflects dispersion)

Example normal distribution

Figure 4: Two Normal p.d.f.s with same variance, but different means

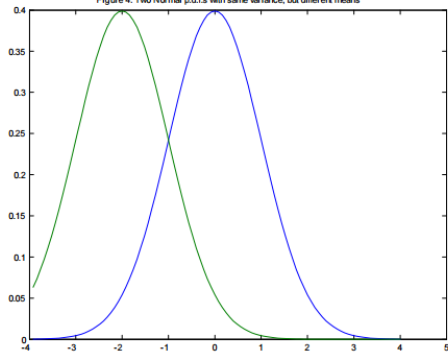
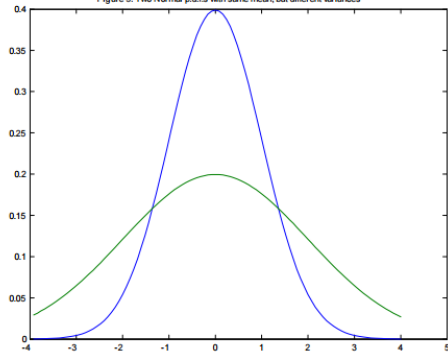


Figure 5: Two Normal p.d.f.s with same mean, but different variances



Standard normal distribution

Standard normal distribution. Properties:

- $N(0,1)$
- Typically the variable is denoted Z and the standard normal cumulative distribution function is denoted with ϕ and $PR(Z \leq c) = \phi(c)$
- A normal distributed variable can be standardized using: $Z = \frac{X - \mu_x}{\sigma_x}$

Using normal statistical tables

- The table for the standard normal distribution can be found in the appendix in the text book.
- Can use $N(0,1)$ tables to figure out probabilities for the $N(\mu, \sigma^2)$ for any μ and σ . If $Y \sim N(\mu, \sigma)$ then:

$$Z = \frac{X - \mu}{\sigma}$$

is $N(0,1)$.

- This is sometimes called the Z-score.
- For any random variable, if you subtract off its mean and divide by the standard deviation you always get a new random variable with mean zero and variance one.

Other common distributions

Chi-square distribution:

- If X has a Chi-square distribution with k degrees of freedom then $X \sim \chi_k^2$
- The chi-square distribution is defined only for positive values of X .
- Is used for comparing estimated variance values to the values based on theoretical assumptions.

Other common distributions

Student t:

- $X \sim t_k$ - student t with k degrees of freedom
- The student-t is bell-shaped and is symmetric
- Used to calculate confidence intervals (using the critical t-value)

F-distribution:

- $X \sim F_{k_1, k_2}$, F F distribution with k_1 degrees of freedom in the numerator and k_2 degrees of freedom in the denominator
- The F distribution we will use to compute F-tests.
- To save space, F-statistical tables usually only provides the 5% significance statistics.
- F-random variables are always positive.

Reading distribution tables

- The degrees of freedom for student-t and chi-square tells you what row in the statistical table to look at.
- For the F-distribution the degrees of freedom in the numerator tells you the row while the degrees of freedom in the denominator tells you the column to look at.

Bernoulli distribution

- A Bernoulli random variable is a binary random variable, which means that the outcome is either zero or one.
- The Bernoulli distribution of variable G is then:

$$G = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1 - p) \end{cases}$$

- The simplicity of the Bernoulli distribution makes the variance and mean simple to calculate

Actual vs asymptotic distribution

Asymptotic distribution

The **asymptotic distribution** is the approximate sampling distribution of a random variable computed using a large sample. The **exact distribution** or the finite-sample distribution is the sampling distribution that exactly describes the distribution.

Many distributions approximate the normal distribution when the number of observations of the random variable increase. This means that we can use the normal distribution to compute approximate probabilities.

Statistics

Key terms

- Population: The "universe" all items (countries, individuals etc) of interest.
- Sample: A subset (preferably random) of the population.
- Parameter: Summary value of the population.
- Statistic: summary measure of the sample.

Random sampling

- A simple random sample is a subset of observations chosen from a larger set (the population).
- Each observation is chosen randomly and entirely by chance such that each member of the population is equally likely to be included in the sample.
- The sample size is of size n while the population is of size N .
- Simple random sampling ensures that the observations are independently and identically distributed (i.i.d)
 - Same probability distribution
 - Mutually independent: The value of Y_1 provides no information about the value of Y_2

Estimates and population parameters

- We are interested in identifying the population parameters: the characteristics of the population distribution function.
- However, we do not typically observe the full population, but only a sample.
- We thus compute estimates of the parameters using estimators.
- An estimate is the numerical value computed using an estimator on a specific sample.
- An estimator is a mathematical rule used to calculate an estimate.

Estimates are typically denoted with a hat.

Statistics

A random sample can be used to:

- Estimate the population mean
- Test hypotheses about the population mean
- Compute a confidence interval for the population mean

Estimators and properties

Choosing estimator

The best estimator is the one that is as close as possible to the unknown true value. More specifically a 'good' estimator for the population mean is:

- Unbiased: $E(\hat{\mu}_y) - \mu_y = 0$ which means that $E(\hat{\mu}_y) = \mu_y$.
- Consistent: $\hat{\mu}_y \xrightarrow{P} \mu_y$.
- Efficient: $var(\hat{\mu}_y) < var(\tilde{\mu}_y)$ where $\tilde{\cdot}$ denotes another estimator.

Estimating the population mean

Potential estimators of the population mean:

- The sample average (\bar{Y}).
- The value of the first observation (Y_1).
- The sample median.
- The sample mode.
- ...

Sample average

The sample average is:

- Unbiased:

$$E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \mu_y$$

- Consistent - when n is large \bar{Y} is close to μ_y with high probability.
- The most efficient in most cases.

$$\text{Var}(\bar{Y}) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{\sigma_y^2}{n}$$

BLUE

The sample average is the Best Linear Unbiased Estimator for the population mean. It is the most efficient among all unbiased estimators that are linear functions of Y_1, \dots, Y_n .

Proof BLUE - Unbiased

The sample average is a unbiased estimator of the population parameter:

$$E(\bar{Y}) = \mu_y$$

The expected value of the sample average is the true population parameter. Thus the sample average is unbiased.

Rules of summation:

- $\sum aX = a \sum X$
- $E(X + Y) = E(X) + E(Y)$
- $\sum a = na$

Proof BLUE - Consistent

The sample mean is consistent if the probability that \bar{Y} is in the range $(\mu_y - c)$ to $(\mu_y + c)$ becomes arbitrarily close to 1 as n increases for any constant $c > 0$. (Key Concept 2.6)

Law of large numbers

Under general conditions, the sample average will be close to the population mean with very high probability when the sample is large.

Proof BLUE - Efficient

The sample average vs Y_1

- Both are unbiased: $E(Y_1) = \mu_y = E(\bar{Y})$
- Both are consistent under the same assumptions
- $Var(Y_1) = \sigma_y^2 > Var(\bar{Y}) = \frac{\sigma_y^2}{n}$

The sample average is a more efficient estimator than using the first observation.

Sampling distribution of the sample average

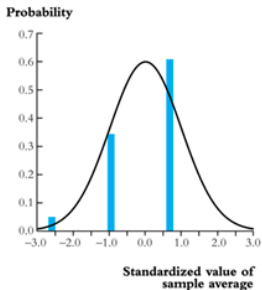
For small sample sizes the distribution of \bar{Y} is complicated but if n is large the central limit theorem dictates that it is asymptotically normal:

Central limit theorem

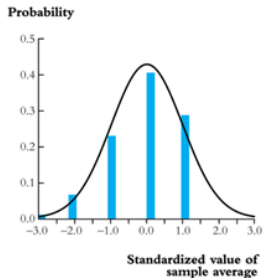
Under general conditions when the sample size is large: the sampling distribution of the standardized sample average is well approximated by a standard normal distribution.

That is: \bar{Y} is approximately $N(\mu_y, \sigma_{\bar{Y}}^2)$

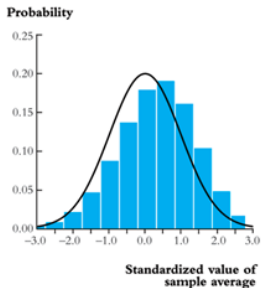
Example: CLT



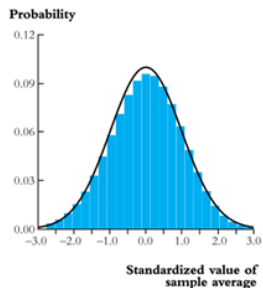
(a) $n = 2$



(b) $n = 5$



(c) $n = 25$



(d) $n = 100$

Non-random sampling

Non-random sampling

Non-random sampling may result in unbiased, inconsistent, estimators even if the estimator are unbiased under random sampling.

Example

If we want to estimate the height of all University of Oslo students where should we go to sample?

- Bus stop?
- Student union?
- Students visiting office during office hours?

Are there any problems with these sampling points?

Measure of distribution

Standard error

$$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = \frac{s_y}{\sqrt{n}}$$

- The standard error of the sample average says something about the uncertainty around the estimate of the mean.
- It is an estimate of how far the sample mean is likely to be from the population mean.
- The standard error falls as the sample size increases as the extent of chance variation is reduced.
- The standard error is used to indicate uncertainty around the estimate.

Measure of distribution:

Sample standard deviation

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- The sample standard deviation is the degree to which individuals within the sample differs from the sample mean.
- The sample standard deviation will tend to the population standard deviation (σ_y) as the sample size increases.
- The sample standard deviation is used to describe how widely scattered the measurements are.

Confidence intervals

Confidence intervals and hypothesis testing

- The confidence interval is an interval estimate for the population parameter such that the chance that the true population parameter lies within that interval is $1 - \alpha$.
- For example

$$Pr(t_{\alpha/2, \nu}) < z < Pr(t_{(1-\alpha/2), \nu}) = 1 - \alpha$$

Confidence interval mean

- A 95% confidence interval for μ_y is an interval that contains the true value of μ_y in 95% of repeated samples.
- 95% confidence interval: $\mu_y = \{\bar{Y} + / - 1.96SE(\bar{Y})\}$

Confidence interval in Stata

```
. mean popgrowth
```

```
Mean estimation                Number of obs   =           68
```

	Mean	Std. Err.	[95% Conf. Interval]	
popgrowth	.9720588	.1129236	.7466624	1.197455

The 95% confidence interval for the mean is automatically reported. You can get another confidence interval by adding the option for level to the command.

```
. mean popgrowth, level(90)
```

```
Mean estimation                Number of obs   =           68
```

	Mean	Std. Err.	[90% Conf. Interval]	
popgrowth	.9720588	.1129236	.7837118	1.160406

Hypothesis testing

Hypothesis testing

Potential questions:

- Does the increase of Co₂ concentration increase the average temperature?
- Are females discriminated against in hiring decisions?
- Are taller people paid more?

The method for answering such questions using a sample of data is known as hypothesis testing.

Hypothesis (1)

- An hypothesis is a statement about the true value of a parameter of interest.
- For example it can be a conjecture about the true value of the population average of Y .
- The statement to be tested is typically called the null hypothesis and is denoted H_0
- The hypothesis against which the null is tested is called the alternative hypothesis typically denoted H_1 or H_A and should hold whenever the null doesn't hold.
- We use data to compare the null hypothesis to the alternative hypothesis.

Example hypothesis

The null hypothesis about the population mean is that it is $\mu_{Y,0}$, thus we can write it as:

$$H_0 : E(Y) = \mu_{Y,0}$$

We can formulate three alternative hypotheses:

- The true value is larger than the null hypothesis value (1-sided.)
- The true value is smaller than the null hypothesis value (1-sided.)
- The true value is different (smaller or larger) from the null hypothesis value (2-sided.)

Important:

Rejecting the null hypothesis does not imply accepting the alternative.

Hypothesis (2)

In hypothesis testing we can make two kinds of mistakes.





Errors in statistical hypothesis tests

Type I error: Rejecting the null hypothesis when it is true.

Type II error: Not rejecting the null hypothesis when it is false.

- The probability of a Type I error is called the significance level, usually denoted by α
- The power of the test is one minus the probability of a Type II error.

Classical hypothesis testing requires that we initially specify a significance level for a test and then maximize the power of the test against the relevant alternatives.

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
R e s e a r c h	The Null Hypothesis Is True	Accurate $1 - \alpha$ 	Type II Error β 
	The Alternative Hypothesis is True	Type I Error α 	Accurate $1 - \beta$ 

Significance level

Significance level

The prespecified rejection probability of a statistical hypothesis test when the null hypothesis is true.

Significance probability

The probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming that the null hypothesis is correct.

Power

Power

The probability that a test correctly rejects the null hypothesis when the alternative is true.

- Power = $1 - P(\text{type II error})$

Example (2)

- To evaluate the hypothesis about the population mean the sample average is compared to the hypothesized value.
- The sample average (\bar{Y}) can differ from the hypothesized value ($\mu_{y,0}$) either because:
 - The null hypothesis is false, i.e. the true mean does not equal $\mu_{y,0}$
 - Random sampling, i.e. the true mean does equal $\mu_{y,0}$, but the sample average differs from the true value due to the nature of random sampling.

Notation

	Estimated from the sample	Parameter from the population
Mean	$\hat{\mu}_x$	μ_x
Variance	$\hat{\sigma}_x^2$	σ_x^2
Standard Deviation	$\hat{\sigma}_x$	σ_x
Size	n	N

- \hat{Y} is the estimate of the parameter Y

Notation

Standard deviation and standard error

- (σ) - Population standard deviation, the true standard deviation in the population.
- Sample standard deviation: (s) An estimate of the population standard deviation.
- s_y is the estimate for population standard deviation for the random variable Y .
- Standard error of an estimator: An estimate of the standard deviation of the estimator.
- $SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = s_y/\sqrt{n}$ is the standard error of sample mean, which is an estimate of the standard deviation of \bar{Y} .
- The sample mean is an estimator of the population mean.

Notation in the book

Population parameter	Sample statistic
μ = Population mean	\bar{Y} = Sample estimate of population mean
σ = Population standard deviation	s = sample standard deviation, estimator of σ
$\sigma_{\bar{y}}$ = Standard deviation of \bar{Y}	$SE(\bar{Y}) = \hat{\sigma}_{\bar{y}}$ = Standard error of \bar{Y} estimator of $\sigma_{\bar{y}}$.

Summary

- Reviewed different types of distributions where the area under the p.d.f gives you probabilities.
- Learned to use statistical tables to obtain these probabilities (seminar 1)
- The normal, chi-square, student-t and F-distributions are the main distributions in this course.
- Learned about confidence intervals and hypothesis testing.
- Looked at the notation in this course.

Exercises

- Assume that the null hypothesis is chi-square distributed under the null hypothesis with 30 degrees of freedom. The relevant test statistic is 40. Can we reject the null at 5% significance?
- Assume that the null hypothesis is t-distributed with 25 degrees of freedom. With a test statistic of 3.0. Can you at 1% significance reject H_0 ?
- Assume that the distribution under the null hypothesis is $F_{4,60}$ and the test statistic is 5.0. Can you reject the null at 5% significance?

Exercises

Assume that heights of women are distributed normally with mean 168 and a standard deviation of 6 cm. What is the probability that a woman is between 160 and 170 cm tall?