

Statistics for Data Science

Lecture 2

Dennis Fok (Econometric Institute)

September – October, 2025

Before next time

Assignment for next week

- Read
 - Chapter 1 (this week's material)
 - Chapter 2 (next week)
- Try some examples in the book yourself (see [here](#) for data and code)
- Finish today's assignments (1.1 - 1.4)
- Continue to practice using own data (or the housing data)
 - Create simple plots
 - Calculate summary statistics
 - Inspect distributions of some variables (also consider transformations of variables)
 - Visualize relations between variables
- Optional: Exercise 2 (Volkswagen prices)

Today

- Distributions
- Getting ready for statistical hypothesis testing...
 - Quantifying uncertainty (standard errors)
 - Confidence intervals

Distribution functions

Some theory on distributions

- Distributions describe the probabilities of events related to random variables.
- Many "standard" distributions exist
 - Bernoulli
 - Binomial
 - Negative Binomial
 - Normal
 - Poisson
 - ...
- What is a distribution?
 - **Mathematical functions** to summarize "probabilities" of events

We need to distinguish between

- Discrete random variables
- Continuous random variables



Slide 4 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Distributions for discrete random variables

Consider a discrete random variable X

- outcomes in **set**: eg. $\{0, 1, 2, \dots, K\}$
(K may be infinite)
- **cumulative distribution function** (often written as $F(x)$ or $P(x)$):
function that gives probability that $X \leq x$ for any x
- **probability mass function**: (often written as $f(x)$ or $p(x)$)
function that gives probability that $X = x$ for any x

Note

- X is the random variable
- x is a particular value/outcome



Slide 5 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Example: tossing a dice

x=outcome	prob. mass function	cumulative distr. function
0	0	0
1	1/6	1/6
2	1/6	2/6
⋮	⋮	⋮
6	1/6	1

$$\text{Prob. mass function} = \Pr[X = x] = f(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Cum. distr. function} = \Pr[X \leq x] = F(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{6} & \text{if } 1 \leq x < 2 \\ \frac{2}{6} & \text{if } 2 \leq x < 3 \\ \vdots & \vdots \\ 1 & \text{if } 6 \leq x \end{cases}$$

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Distributions for continuous random variables

Consider a continuous random variable X

- outcomes in interval (a, b)
(a may be $-\infty$ and/or b may be $+\infty$)
- **cumulative distribution function (cdf)**:
function $F(x)$ that gives probability that $X \leq x$ for any x
(same as before)
- Probability that X equals x (eg. $x = 1.335221$)?
→ this equals 0 exactly!
- **probability density function (pdf)**:
function $f(x)$ such that areas under the curve correspond to probabilities

$$\Pr[a < X \leq b] = \int_a^b f(x) dx = F(b) - F(a)$$

→ Note: $f(x)$ is the derivative of $F(x)$



Slide 7 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Interpretation probability mass/density function

Consider the function at some value x :

- Discrete variables:
 - Clear interpretation
 - Probability of outcome x
- Continuous variables:
 - No direct formal interpretation
 - "Indication of relative frequency of outcomes *close to* x "
 - Area under curve gives probability

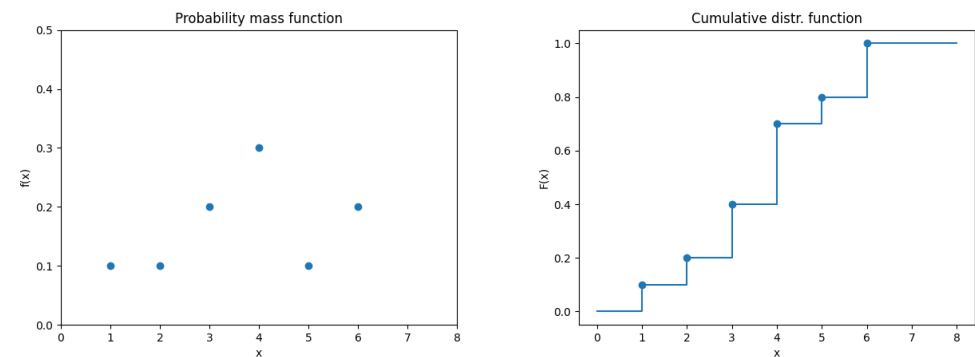
Erasmus

Slide 8 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Graphical illustration – Discrete example

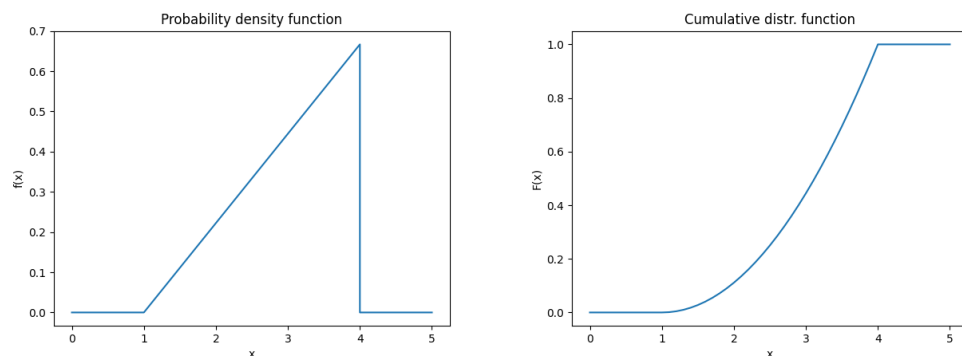
Discrete random variable



© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Graphical illustration – Continuous example

Continuous random variable



© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Graphical illustration – Continuous example

For the previous example we have

$$f(x) = \begin{cases} \frac{2}{9}(x-1) & \text{if } 1 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

and


$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{9}(x^2 - 2x + 1) & \text{if } 1 \leq x \leq 4 \\ 1 & \text{if } x > 4 \end{cases}$$




Question/Assignment (Advanced/For fun/At home)

Check that $f(x)$ and $F(x)$ indeed match and that both satisfy the requirements for a density/distribution function






© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Distribution functions in Python

 `scipy.stats` contains many many distributions

- Normal ( `norm(..)`)
- Binomial ( `binom(..)`)
→ number of successes in k trials, if each trial has success probability p
- Exponential ( `expon(..)`)
- ...

Standard available functions

- probability density function (method  `.pdf()`)
- probability mass function (method  `.pmf()`)
- cumulative distribution function (method  `.cdf()`)
- quantile function (inverse cdf) (method  `.ppf()`=percent point function)
- generate random numbers (method  `.rvs()`)




Slide 12 of 45


Examples:

Calculate cdf for standard normal distribution at 1:

 `scipy.stats.norm.cdf(1, loc=0, scale=1)`

or

 `scipy.stats.norm(loc=0, scale=1).cdf(1)` (this makes clear which parameters belong to which part)

If you use  `from scipy.stats import norm` you can omit the `scipy.stats` part





Slide 13 of 45

Assignment

In-class assignment 2.1

- 1 Calculate the probability that...
 - a standard normally distributed variable is larger than 1.
 - a normally distributed variable with mean 20 and variance 10 is smaller than 15.
 - we get (exactly) 15 times head in 30 coin tosses.

Hint: use for example  `help(norm.cdf)` and  `help(binom.pmf)` to find out how to use these functions
- 2 Suppose that a soccer club has a 60% probability of winning each match they play. What is the probability that they do **not** win any of the first four matches of the year?
- 3 Calculate a quantile
 - Suppose that the waiting time for the bus has an exponential distribution with scale 10. How many minutes does one have to wait at least on the 5% worst days?

→ Use `inclass_2.1.py` (on Canvas) to get started



Slide 14 of 45

Special distributions

Standard distributions

We will often use four special distributions

- Normal distribution: $X \sim N(\mu, \sigma^2)$ (\sim means “has distribution”)
- F distribution: $X \sim F(d_1, d_2)$
- Chi-squared distribution: $X \sim \chi^2(k)$
- t distribution: $X \sim t_n$

(these 4 distributions are strongly related)

In Python use

- `norm(μ, σ)`, `f(d_1, d_2)`, `chi2(k)`, `t(n)` from `scipy.stats`

→ see also [scipy documentation](#), wikipedia and [the distribution zoo](#) for more information on distributions (including formulas and references)

Erasmus

Slide 15 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Normal distribution

Normal distribution with mean μ and variance σ^2 : $N(\mu, \sigma^2)$

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

`scipy.stats.norm(loc=mu, scale=sigma).pdf(x)` and

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(z) dz \\ &= \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(z - \mu)^2}{\sigma^2}\right) dz \end{aligned}$$

`scipy.stats.norm(loc=mu, scale=sigma).cdf(x)`

If $\mu = 0$, $\sigma = 1$

- standard normal
- Usual notation: $f(x) = \phi(x)$ and $F(x) = \Phi(x)$

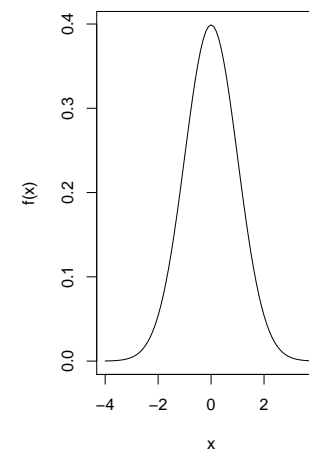
Erasmus

Slide 16 of 45

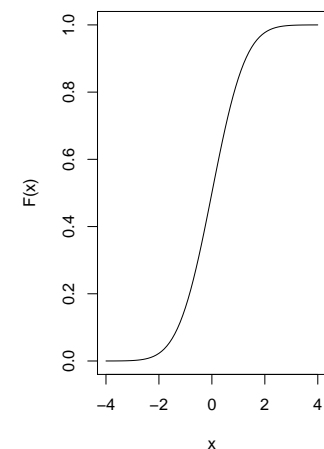
© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Standard normal

Probability density function



Cumulative distr. function



© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Useful properties of the normal

Given $X \sim N(\mu, \sigma^2)$ and a, b numbers (not random)

- $a + X \sim N(a + \mu, \sigma^2)$
- $bX \sim N(b\mu, b^2\sigma^2)$
- $a + bX \sim N(a + b\mu, b^2\sigma^2)$
- $\frac{X - \mu}{\sigma} \sim N(0, 1)$

Given $Y \sim N(\alpha, \nu^2)$ independent of X

- $X + Y \sim N(\mu + \alpha, \sigma^2 + \nu^2)$



Slide 18 of 45

Definitions of distributions related to the normal

Suppose

- $X_1, \dots, X_k \sim N(0, 1)$
- $Y_1, \dots, Y_n \sim N(0, 1)$
- All X_i and Y_j independent

We have

- For “sums of squares”: $\sum_{i=1}^k X_i^2 \sim \chi^2(k)$ and $\sum_{i=1}^n Y_i^2 \sim \chi^2(n)$
- Ratio of “mean squares”

$$\frac{\sum_{i=1}^k X_i^2 / k}{\sum_{i=1}^n Y_i^2 / n} \sim F(k, n)$$

- Ratio of normal to “root mean squares”

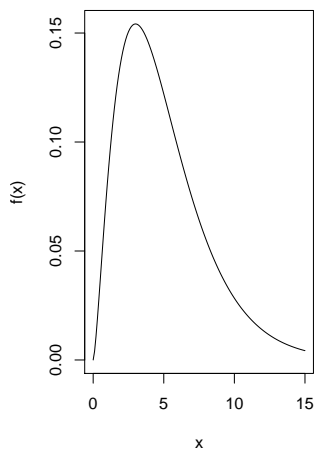
$$\frac{X_i}{\sqrt{\sum_{i=1}^n Y_i^2 / n}} \sim t_n$$



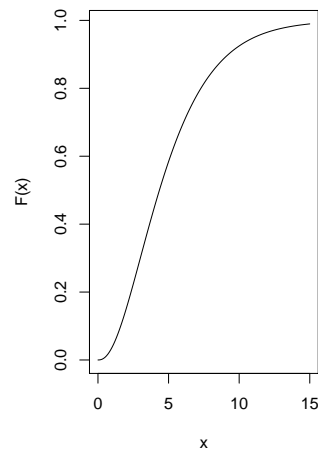
Slide 19 of 45

Chi-squared(5)

Probability density function

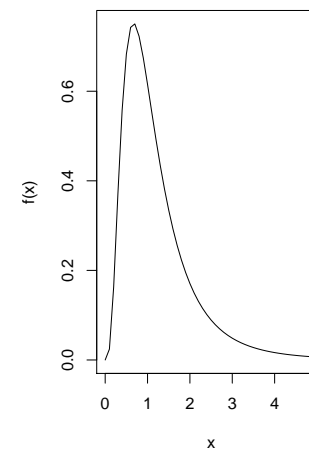


Cumulative distr. function

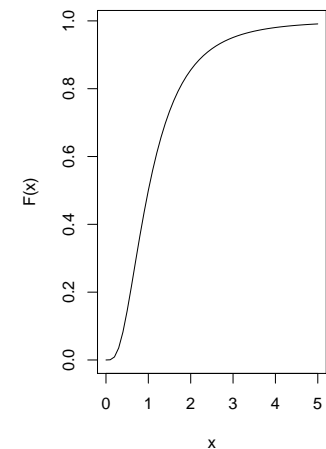


F(10,10)

Probability density function



Cumulative distr. function



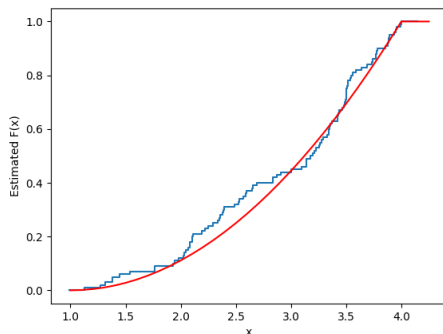
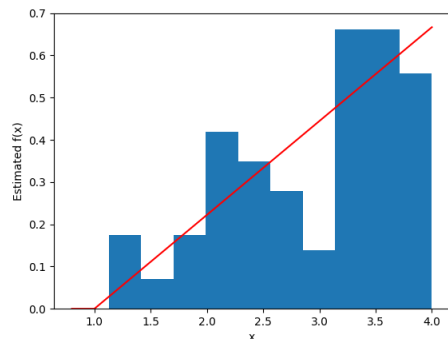
Estimating distributions

Histograms vs. densities

- Histogram → (discretized) estimated density (up to scaling)
 - Histogram = estimate
 - Density = population equivalent (often unknown)
- Continuous *density* estimators also exist (have seen this already)
- Can also estimate cumulative distribution function
 - Plot sorted data vs. index/n
→ `scipy.stats.ecdf(data).cdf.plot()`
 - Use the `scipy.stats.fit()` functions to fit a specific distribution

Histogram vs. true density (in red)

Given 100 observations:



Fitting distributions

The package `scipy.stats` can be used for

- 1 Plotting the empirical distribution
- 2 Fitting a *particular* distribution to data
- 3 Inspecting the fit
- 4 Comparing the fit of different distributions

Key functions

- `fit = stats.fit(stats.norm, data, bounds)`
→ fit a normal distribution to the data (or different distributions), parameters are within specified bounds → choose these sensibly based on your data.
For example `bounds = {'loc': (-4,4), 'scale': (0,1)}`
- `fit.plot()` and `fit.plot(plottype=t)` with `t` one of "hist", "qq", "pp", "cdf"
- `fit.nllf()`
→ Negative log-likelihood → measures fit (lower=better)

Assignment

In-class assignment 2.2

- 1 Look at the example code in `inclass_2.2.py`
- 2 Load the houseprice data (from week 1)
- 3 Use `.hist()` to show the empirical density of lotsize
- 4 Fit a normal ("norm") and a log-normal ("lognorm") distribution to the lotsize and graphically inspect the fit of both
- 5 Which one fits better on the basis of the graphical inspection?

Quantifying estimation uncertainty

Estimation uncertainty

Last week:

- Moments of random variables (mean, variance, etc)
- Remember that we have
 - 1 theoretical moments: eg. expected value $E[X]$
 - 2 sample moments: eg. sample mean $\frac{1}{n} \sum_i X_i$

Remember:

- Moments are (in principle) unknown
- Given a sample X_1, \dots, X_n :
 - we **estimate** the population mean and
 - **estimate** the population variance!

Question

How big can estimation uncertainty be?

Illustration of estimation uncertainty

Visualizing estimation uncertainty

- Generate some data with known expectation
- Calculate mean
- Compare mean to expectation

→ Repeat many times!

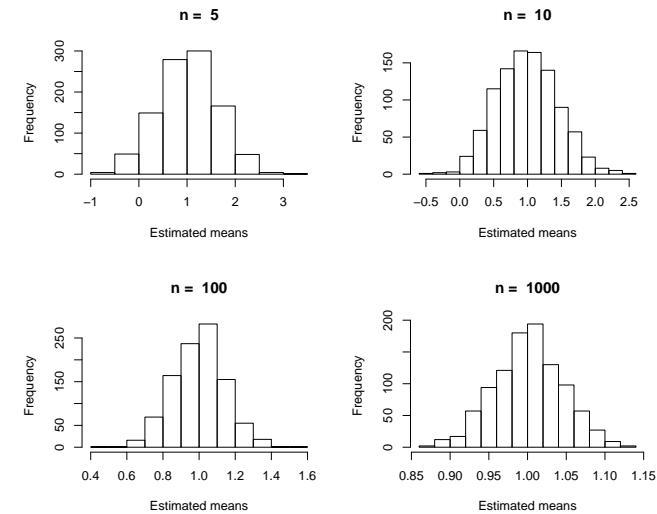
Interesting questions

- How bad can things get?
→ look at min/max
- How bad are things on average?
→ look at variance of found means over data sets

Erasmus

Slide 27 of 45

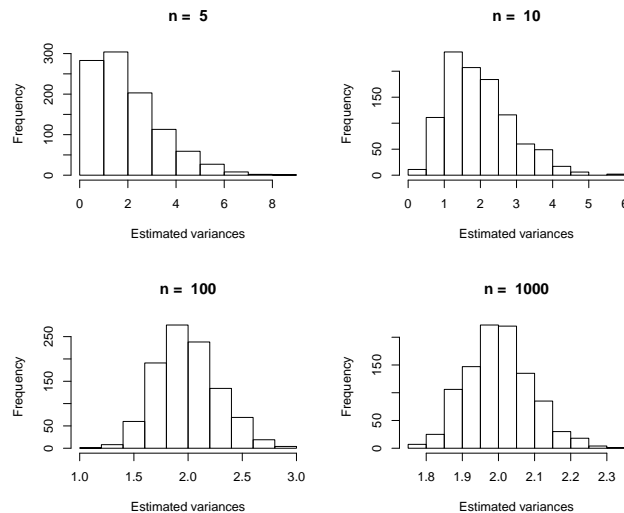
Example with expectation=1, variance=2: Sample mean



Erasmus

Slide 28 of 45

Example with expectation=1, variance=2: Sample variance



Erasmus

Slide 29 of 45

Estimation uncertainty

In statistics there will always be **estimation uncertainty**

→ Quantifying the uncertainty is important!

How uncertain is an estimate of the population mean with n observations?

→ Calculate the variance of the estimator!

Assumptions/notation:

- 1 The n observations are independent
- 2 The n observations all have the same distribution
→ mean & variance are constant
- 3 **Population** mean and variance are called μ and σ^2
- 4 Sample is denoted by X_1, \dots, X_n

1. and 2. are abbreviated as *iid* (=independent and identically distributed)

Erasmus

Slide 30 of 45

Estimation uncertainty

Estimator for population mean: $\bar{X} = \frac{1}{n} \sum_i X_i$

Measure uncertainty by variance

$$\begin{aligned}\text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n} \sum_i X_i\right] = \frac{1}{n^2} \text{Var}\left[\sum_i X_i\right] = \frac{1}{n^2} \sum_i \text{Var}[X_i] \\ &= \frac{1}{n^2} \sum_i \sigma^2 = \frac{n}{n^2} \sigma^2 = \frac{1}{n} \sigma^2\end{aligned}$$

(Q: which assumption is used where?)

- Variance of estimator decreases with factor n
- In practice: σ^2 also unknown!
→ replace σ^2 by estimator s^2 (see last week)



Slide 31 of 45

Standard error of the mean

$$\sqrt{\frac{1}{n} s^2}$$

- **Standard error** of the mean
 - Measure of uncertainty in estimated mean
- Standard error = most important measure of uncertainty

Terminology

- Standard **deviation**:
 - 1 square root of theoretical variance *or*
 - 2 square root of sample variance
- Standard **error**
 - 1 square root of *estimated* variance of an estimator



Slide 32 of 45

Confidence intervals

Confidence intervals

Question:

How to transform the standard error into something interpretable?

→ Random samples come with random variation, what range of variation can we expect?

Assume X_1, \dots, X_n iid from $N(\mu, \sigma^2)$

- $\sum_i X_i \sim N(n\mu, n\sigma^2)$
- $\bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$
- $(\bar{X} - \mu) \sim N(0, \frac{1}{n}\sigma^2)$

If we know σ^2 → we have an idea how large $\bar{X} - \mu$ can be!



Slide 33 of 45

Confidence intervals

Working things out further

$$\frac{(\bar{X} - \mu)}{\sqrt{\frac{1}{n}\sigma^2}} \sim N(0, 1)$$

Therefore

$$\Pr[-1.96 < \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n}\sigma^2}} < 1.96] = 0.95$$

`stats.norm.ppf(0.025)` `stats.norm.ppf(0.975)`

So

$$\Pr[\mu - 1.96\sqrt{\frac{1}{n}\sigma^2} < \bar{X} < \mu + 1.96\sqrt{\frac{1}{n}\sigma^2}] = 0.95$$

Erasmus

Slide 34 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Confidence intervals

Given

$$\Pr[\mu - 1.96\sqrt{\frac{1}{n}\sigma^2} < \bar{X} < \mu + 1.96\sqrt{\frac{1}{n}\sigma^2}] = 0.95$$

Knowing μ and σ^2 :

sample mean falls in the interval $\mu \pm 1.96\sqrt{\frac{1}{n}\sigma^2}$ with probability 95%

Random

Non-random

→ Is this useful in practice?

Erasmus

Slide 35 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Confidence intervals

Without knowing $\mu \rightarrow \mu$ and \bar{X} swap places

- the interval $\bar{X} \pm 1.96\sqrt{\frac{1}{n}\sigma^2}$ contains true value μ with probability 95%

Note: the interval is now the random variable!

What if:

- σ^2 not known
- X_i not normally distributed

Erasmus

Slide 36 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Challenge I: Unknown σ^2

Usually variance σ^2 is not known!

→ replace σ^2 by estimate s^2

Consequences

- $\frac{(\bar{X} - \mu)}{\sqrt{\frac{1}{n}s^2}}$ in general is NOT standard normal
- In fact

$$\frac{(\bar{X} - \mu)}{\sqrt{\frac{1}{n}s^2}} \sim t_{n-1}$$

→ t_{n-k} in general (if k parameters are estimated)

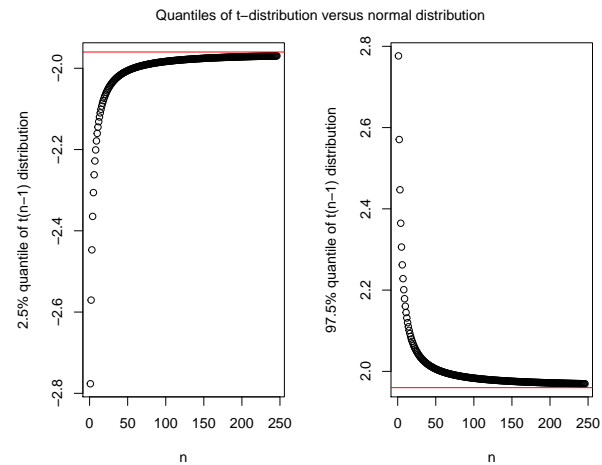
- For large n : $t_{n-k} \approx N(0, 1)$

Erasmus

Slide 37 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Quantiles t_{n-1} -distribution for different n



© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Confidence interval when σ^2 is unknown

- Now use $\frac{(\bar{X} - \mu)}{\sqrt{\frac{1}{n} s^2}} \sim t_{n-k}$:

- Interval

$$\bar{X} \pm t_{n-k}^{0.975} \sqrt{\frac{1}{n} s^2}$$

contains μ with probability 95%

- Note $t_{n-k}^{0.025} = -t_{n-k}^{0.975}$ (Use: `stats.t(n-k).ppf(.)`)

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Erasmus

Slide 39 of 45

Assignment

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

In-class assignment 2.3

- 1 Load the houseprice data
- 2 Use `mean` to get the mean lotsize
- 3 Calculate the mean lotsize directly using the right function
- 4 Calculate the standard error of the mean (see slide 32)
- 5 Use the formula on slide 39 to calculate a 95% confidence interval around the mean (You should obtain the interval [4967.998, 5332.533])

Erasmus

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Slide 40 of 45

Challenge II: non-normal distribution

The normality assumption is needed to get

$$\frac{(\bar{X} - \mu)}{\sqrt{\frac{1}{n} s^2}} \sim t_{n-k}$$

→ What if X_i not normal?

Solution: use **Central limit theorem**

Central Limit Theorem

If $n \rightarrow \infty$: $\sqrt{n} \times$ sample mean converges in distribution to a normal distribution (some regularity conditions are required)

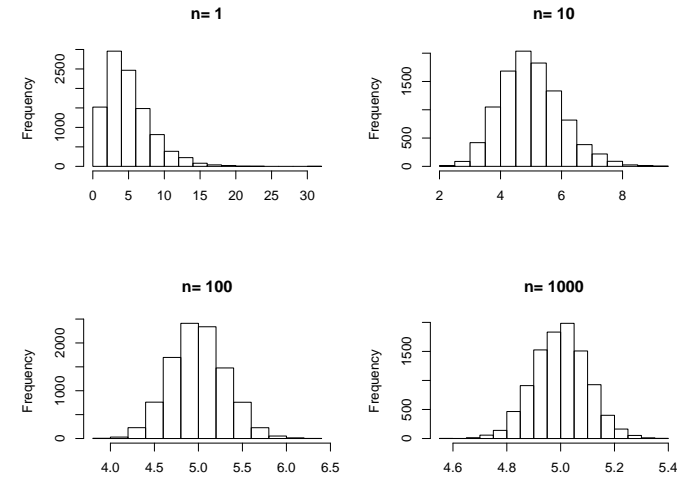
Erasmus

Slide 41 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Illustration CLT

Distribution of sample mean for different n (data from chi-squared(5))



© 2025 Erasmus University Rotterdam. All rights reserved

Use of central limit theorem

Given sample X_1, \dots, X_n and n large enough

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{1}{n} \sigma^2\right)$$

Implication:

- Confidence intervals can be made based on the normal distribution (Note $t_{n-k} \approx N(0, 1)$ for large n)

Erasmus

Slide 43 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

What if n not large?

If n is (too) small **and** data is non-normal

- CLT is not useful
- Confidence intervals based on t-distr. not correct!

Solutions:

- If distribution is known: work out new formulas
- Use non-parametric methods
 - non-parametric tests
 - bootstrap methodology (later lecture)

Erasmus

Slide 44 of 45

© 2025 Erasmus University Rotterdam. All rights reserved. No text and datamining

Before next time

Assignment for next week

- Finish/Reread Chapter 2
- Read Chapter 3 on testing (skip ANOVA and Multi-Arm Bandits)
- Reconsider/finish the in-class assignments
- Look at examples in book
- You can already start working on the “final” assignment (will be on Canvas early next week)

