DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Advanced Probability Theory and Statistics: Inequalities, Convergence of Random Variables, Confidence Intervals, and Hypothesis Tests

Kim Steenstrup Pedersen

How to prepare for these lectures

You should either prior to the lectures, or just after

- Read the material and take notes
- Watch the extended video lectures and take notes
- Browse the slides

Plan for today

- Probability theory
  - Bounds on expectations and tail probabilities
  - Limit theorems for random variables:
    - Weak law of large numbers
    - Central limit theorem
  - Student-t distribution and Chi-square distribution (see video lecture)

- Statistics
  - Confidence intervals
  - Hypothesis tests – the t-test

Computing expectations and probabilities

- It is not always possible to compute expectations and probabilities analytically (i.e. exactly).
- Instead we can do
  - **Bounds using inequalities.** Has many applications in statistics and in theoretical machine learning. This is a topic for this lecture.
  - **Approximations using limit theorems.** This is also a topic for this lecture.
  - **Simulation by sampling** – the Monte Carlo approach. More on this after Christmas.

Inequalities:

Bounds on expectations and tail probabilities

Reading material: Blitzstein & Hwang, Ch. 10.1

Cauchy-Schwarz: A marginal bound on joint expectation

- **Theorem Cauchy-Schwarz:** For any random variables (r.v.) X and Y with finite variances (i.e. $\mathrm{Var}[X] < \infty$ etc),

$E_{P(X)}[X^2]$

$$\left|E[XY]\right| \le \sqrt{E\left[X^2\right]E\left[Y^2\right]}$$

$E_{P(Y)}[Y^2]$

$E_{P(X,Y)}[XY]$

where $|\cdot|$ denotes absolute value.

Meaning: Expectation over the joint distribution is bounded by the marginal second moment expectations. See book for proof.

- **Simple example:** Using the trick $X = X \cdot 1$ and Cauchy-Schwarz (then Y=1), we have $\left|E[X\cdot 1]\right| \le \sqrt{E\left[X^2\right]E\left[1^2\right]}$

Rearranging and substitution gives

$$\left|E[X\cdot 1]\right| = \left|E[X]\right| \le \sqrt{E\left[X^2\right]} \Rightarrow \left(E[X]\right)^2 \le E\left[X^2\right]$$

Hence variance is always nonnegative. $\Rightarrow\ 0 \le E[X^2] - (E[X])^2 = \mathrm{Var}[X]$

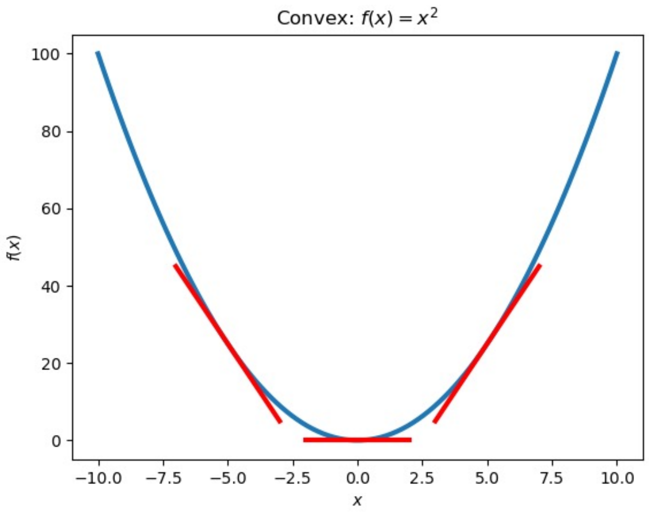
Jensen's Inequality: Functions of r.v.'s and expectations

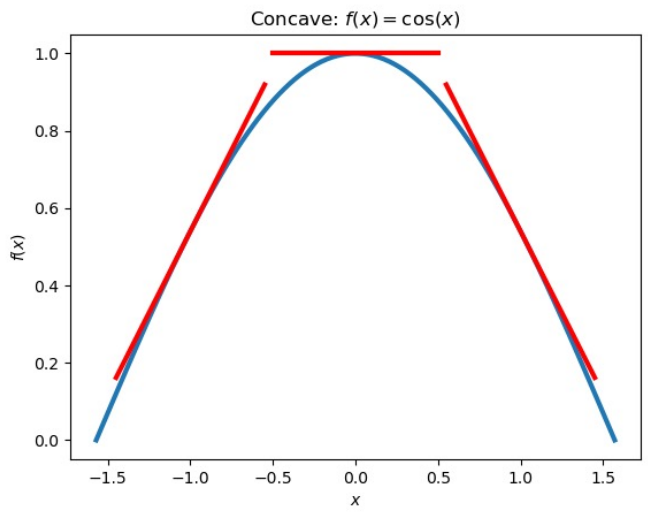
- **Theorem Jensen's Inequality:** Let *X* be a r.v. If *g* is a convex function, then $E\left[g(X)\right] \geq g\left(E[X]\right)$. If *g* is concave, then $E\left[g(X)\right] \leq g\left(E[X]\right)$. Equality holds only, if there are constants *a* and *b*, such that $g(X) = a + bX$, i.e. g is linear (with probability 1).
- Allows us to move expectations in and out of functions.
- Examples:
  - For $g(x) = x^2$ (convex), we get $E[X^2] \geq (E[X])^2$ (known from Cauchy-Schwarz)
  - For $g(x) = |x|$ (convex), we get $E[|X|] \geq |E[X]|$
  - For $g(x) = \log x$ (concave), we get $E[\log X] \leq \log E[X]$ for $X > 0$
  - For $g(x) = a + bx$ (linear), we get $E[g(X)] = E[a + bX] = a + bE[X] = g(E[X])$ (using linearity property of expectation)

Convex and concave functions

Derivative $f'(x) = df/dx$

- **Def.:** A function f differentiable in (a,b) is convex if $f(x_2) \geq f(x_1) + f'(x_1)(x_2 - x_1), \forall x_1, x_2 \in (a,b), x_1 \neq x_2$ and concave if $f(x_2) \leq f(x_1) + f'(x_1)(x_2 - x_1), \forall x_1, x_2 \in (a,b), x_1 \neq x_2$
- Geometric interpretation: The graph of a convex function f never lies below any of its tangents (as given by the derivative f'). The opposite holds for concave functions.





Markov, Chebyshev: Bounds on tail probabilities

- **Theorem Markov:** For any r.v. *X* and constant *a* > 0,

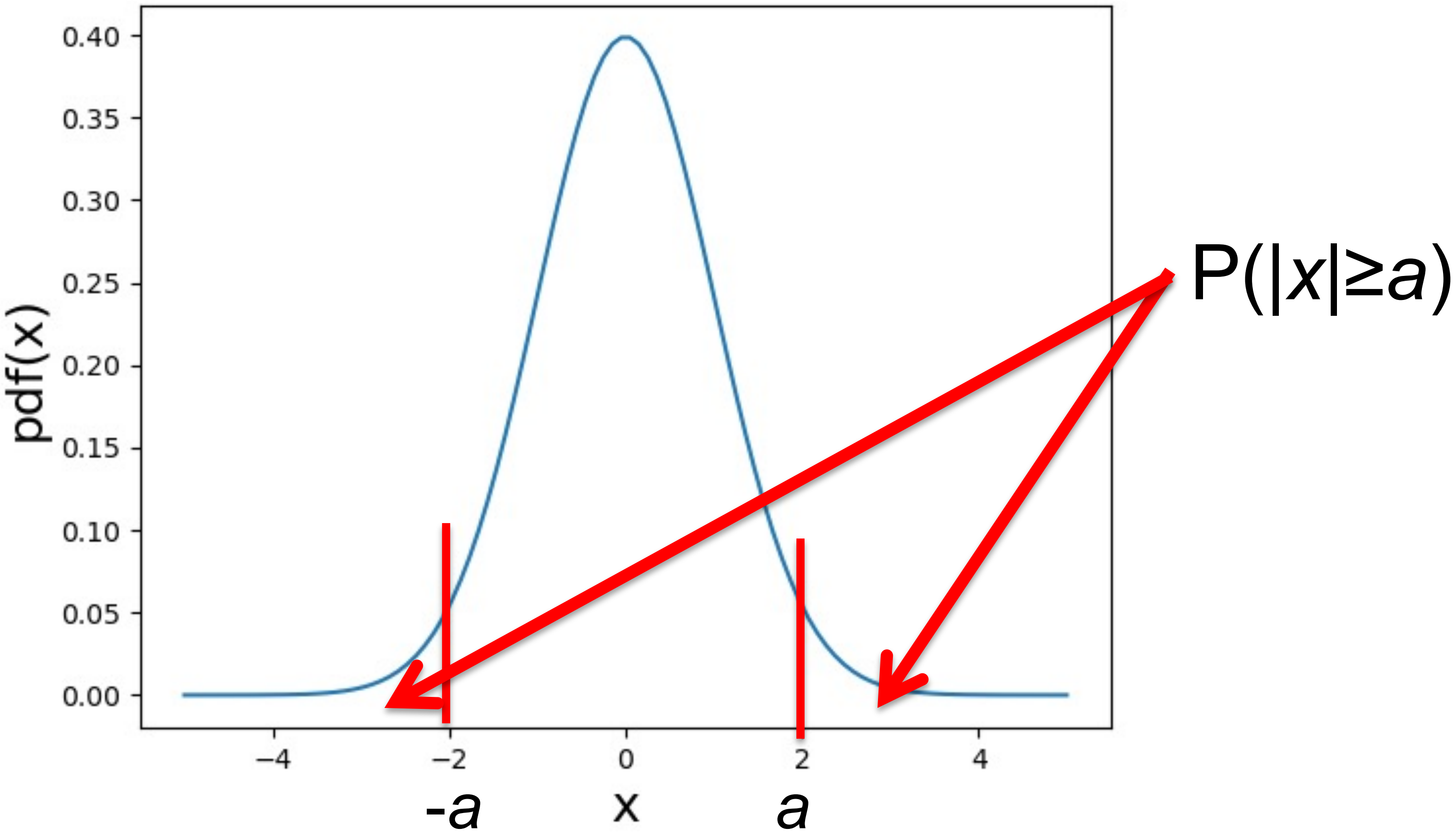
$$P\left(\left|X\right| \geq a\right) \leq \frac{E\left[\left|X\right|\right]}{a}$$

Tail probabilities

$$P\left(\left|X\right| \geq a\right) \leq \frac{E\left[\left|X\right|\right]}{a}$$

Markov inequality for a Normal distributed r.v. *X* with N(0,σ^2)

$$\lim_{a\to\infty} \frac{E\left[\left|X\right|\right]}{a} = 0, \text{ hence } \lim_{a\to\infty} P\left(\left|X\right| \geq a\right) = 0$$



Markov, Chebyshev: Bounds on tail probabilities

- **Theorem Markov:** For any r.v. X and constant $a > 0$,

$$P\left(\left|X\right| \geq a\right) \leq \frac{E\left[\left|X\right|\right]}{a}$$

- **Theorem Chebyshev:** Let X have mean μ and variance σ^2, then for any $a > 0$,

$$P\left(\left|X - \mu\right| \geq a\right) \leq \frac{\sigma^2}{a^2}$$

(A specialization of the Markov inequality)

Limit Theorems:

Convergence properties of sums of random variables

Reading material: Blitzstein & Hwang, Ch. 10.2 – 10.3

Or

Pishro-Nik, Ch. 7.0 – 7.1

Sample mean

- Consider **independent and identically distributed (i.i.d.)** r.v.'s $X_1, X_2, X_3, \ldots$ with finite mean μ and finite variance σ^2 and the sample mean

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

- Realize that this is also a r.v. (a function of r.v.'s) with

$$E\left[\bar{X}_n\right] = \frac{1}{n} E\left(X_1 + \cdots + X_n\right) = \frac{1}{n}\left(E\left[X_1\right] + \cdots + E\left[X_n\right]\right) = \frac{1}{n}\left(n\mu\right) = \mu$$

$$\mathrm{Var}\left[\bar{X}_n\right] = \frac{1}{n^2} \mathrm{Var}\left(X_1 + \cdots + X_n\right) = \frac{1}{n^2}\left(\mathrm{Var}\left[X_1\right] + \cdots + \mathrm{Var}\left[X_n\right]\right)$$

$$= \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}$$

Variance of sample mean:
$\lim_{n\to\infty} \mathrm{Var}[\overline{X_n}] = 0$

Law of Large Numbers

- **Intuition:** The law of large numbers state that as *n* increases, the sample mean $\bar{X}_n$ converges to the true mean μ. It comes in two flavours – the strong and weak law of large numbers.
- **Theorem Weak law of large numbers:** For all $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) = 0$$

- Proof: We just need to use Chebyshev's inequality

$$P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{n\varepsilon^2}$$

$$\boxed{P\left(\left|X - \mu\right| \ge a\right) \le \frac{\sigma^2}{a^2}}$$

- and since $\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$, so does the probability.

The Central Limit Theorem

- What's the distribution of the sample mean r.v. $\bar{X}_n$ as *n* increases?
- **Central Limit Theorem (CLT):** As $n \to \infty$,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to N(0,1) \text{ in distribution}$$

we consider the distribution of this r.v.

- **Note:** Standardization of r.v. refers to subtracting the mean and division by the standard deviation. This is done above to $\bar{X}_n$

The central limit theorem in practice

- CLT says that the distribution of a sum of random variables converges to a Gaussian distribution, e.g.
$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$
- It does not matter what the distribution of the individual X_i is, as long they are i.i.d. and have finite mean, $0 \leq |\mathrm{E}[X_i]| < \infty$, and finite variance, $0 < \mathrm{Var}[X_i] < \infty$.
- You do not need a large n for CLT to hold. To see this, write a small program that generates samples from a Uniform distribution $U(0,1)$ and compute sample mean 1.000 times for an increasing list of n values (e.g. $n \in \{1, 2, 50, 100\}$). Plot the histogram over the 1.000 estimates for each n value.
- Lets look at the example in `clt.py`

Distributions:

Lets look at a couple of named probability distributions we need now

Reading material: Blitzstein & Hwang, Ch. 10.4

Chi-square (χ_n^2) distribution

Example: Sum of sample variances

- **Definition:** Let $V = Z_1^2 + \ldots + Z_n^2$ where $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$. Then V is said to have the Chi-square distribution with n degrees of freedom and we write $V \sim \chi_n^2$.
- The χ_n^2 distribution is a special case of the Gamma distribution,

$$\text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$$

- The probability density function (PDF) is given by

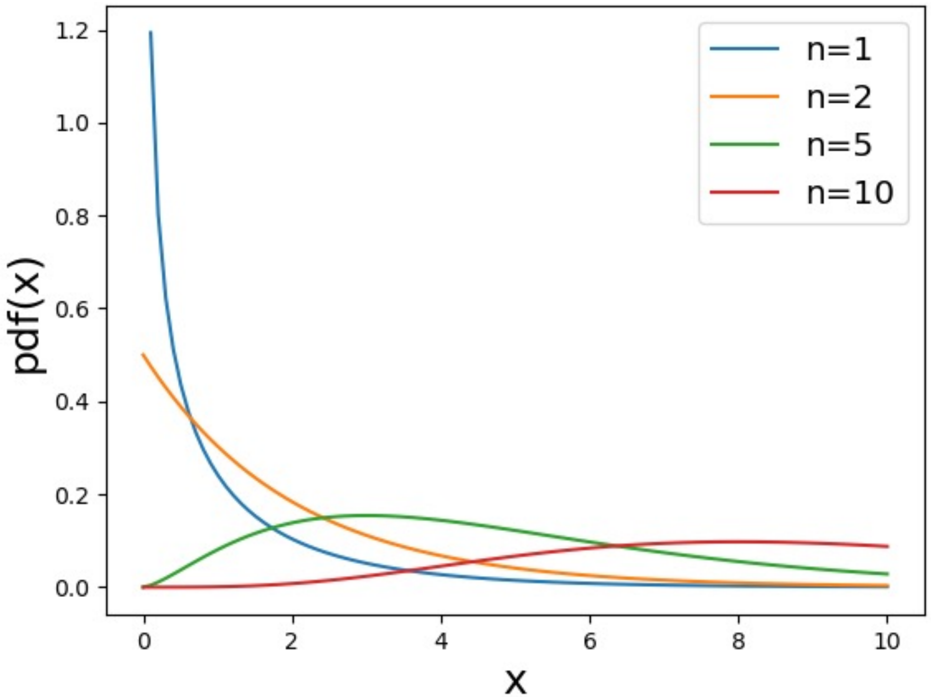
$$f_V(v) = \frac{1}{\Gamma(n/2)}\left(\frac{1}{2}v\right)^{n/2}\frac{1}{v}e^{-\frac{1}{2}v} \quad , \quad v > 0$$

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt, \quad z \in \mathbb{C}$$

Relates to the distribution of sample variance.

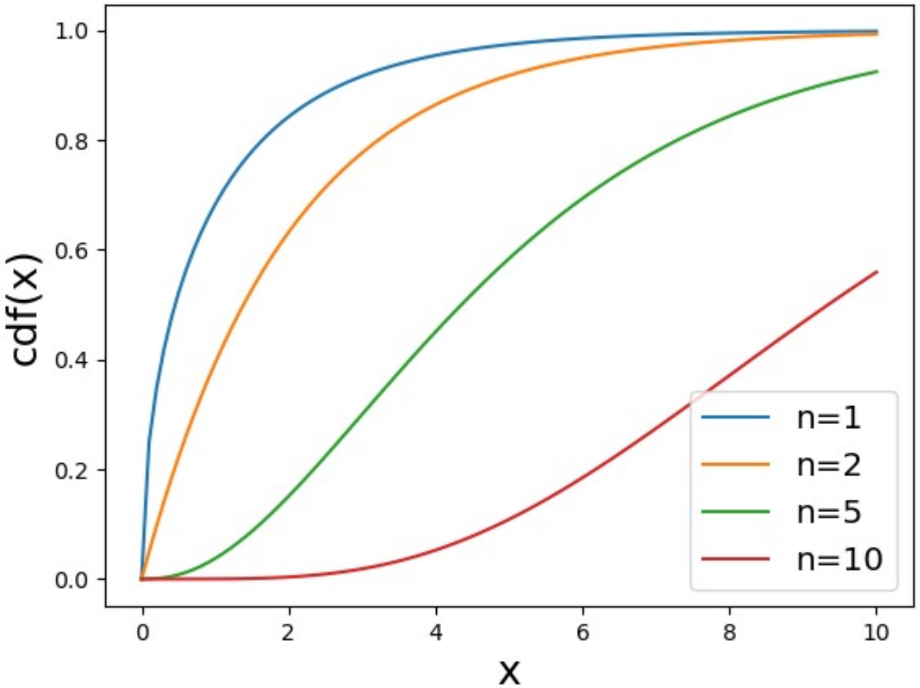
Visualizing the Chi-square (χ_n^2) distribution

PDF (probability density function)



`scipy.stats.chi2.pdf(x,n)`

CDF (cumulative distribution function)



`scipy.stats.chi2.cdf(x,n)`

(Student's) t-distribution

- **Definition:** The t-distribution with *n* degrees of freedom is defined by this r.v.

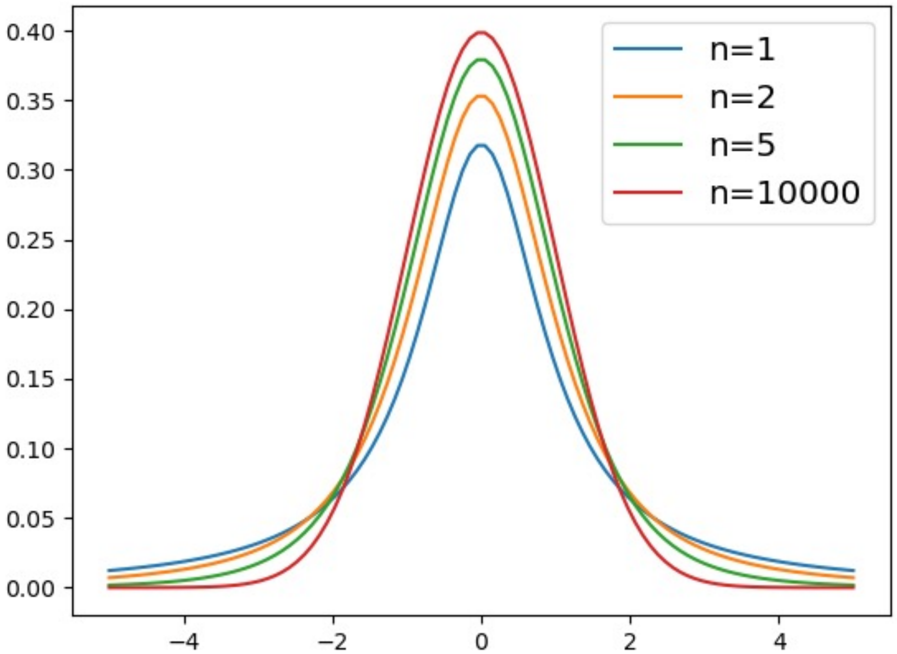
$$T = \frac{Z}{\sqrt{V/n}}$$

  where $Z \sim N(0,1)$ and $V \sim \chi_n^2$ and *Z* is independent of *V*.
- The PDF is given by

$$f_T(t) = \frac{\Gamma\left((n+1)/2\right)}{\sqrt{n\pi}\,\Gamma\left(n/2\right)}\left(1 + t^2/n\right)^{-(n+1)/2}$$

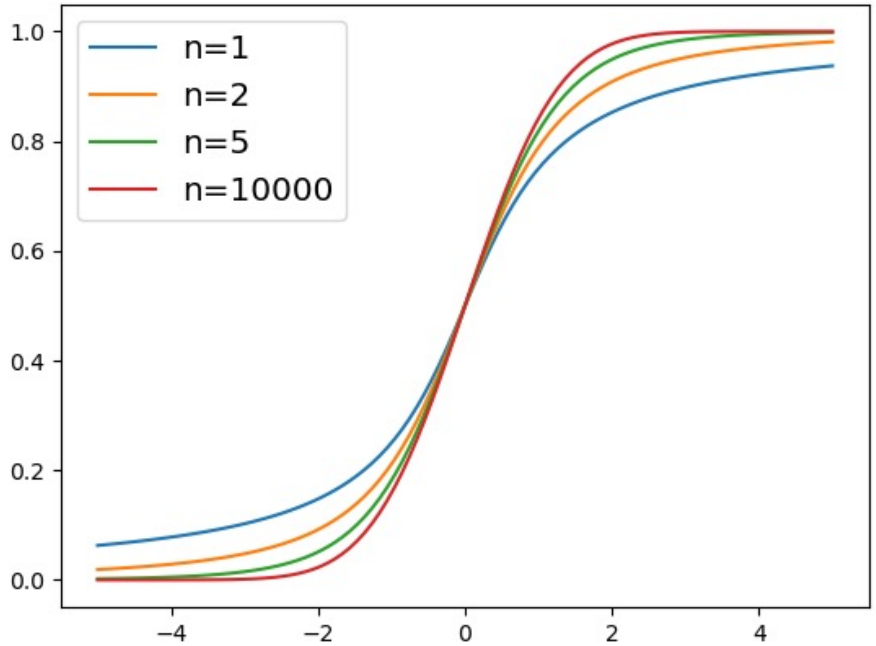
Visualizing the t-distribution

PDF



`scipy.stats.t.pdf(x,n)`

CDF



`scipy.stats.t.cdf(x,n)`

The distribution is symmetric and has two distributions as special instances (Cauchy distribution at n=1 and Normal distribution as $n \to \infty$)

Confidence intervals

Reading material: Kreyszig, Ch. 25.1, 25.3

Or

Pishro-Nik, Ch. 8.1 – 8.2.2, 8.3

Parameter estimation – point estimate

- We want to estimate parameters of a probability distribution model (e.g. the mean and variance in a Normal distribution).
- The function computing the estimate from sampled data is called an **estimator**.
- If the estimator provides a specific value for our parameter, this is called a **point estimate**.
- **Example:** Computing the sample mean of a Normal distributed r.v. is a point estimator for the mean parameter. Let $x_1, \ldots, x_n$ be sampled data from a Normal distributed r.v. X, then the **sample mean** is

$$\bar{x} = \frac{1}{n}\left(x_1 + \ldots + x_n\right)$$

This is deterministic once $x_1, \ldots, x_n$ are fixed

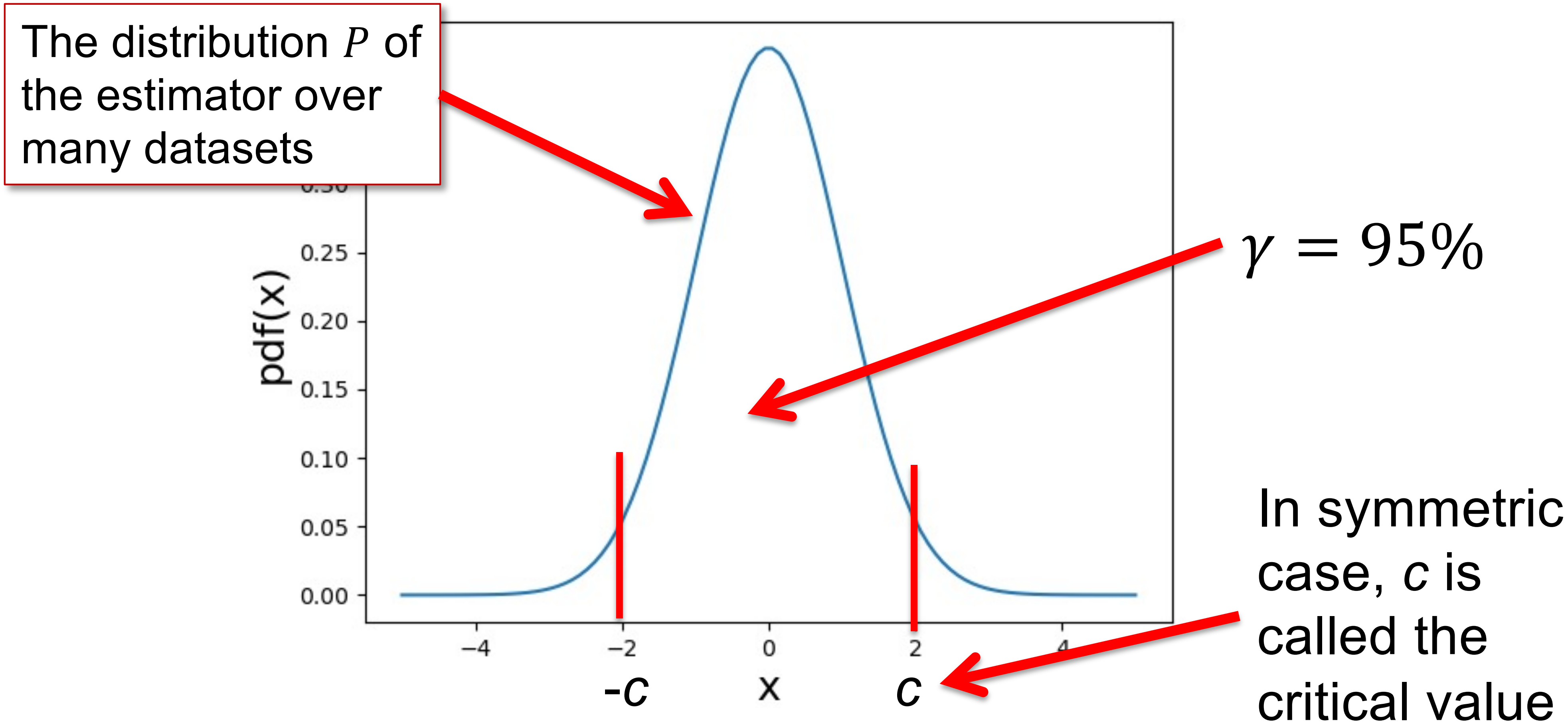
Parameter estimation – confidence interval

- We can also compute an interval in which the true estimate (value) lies with a chosen probability (confidence level). This is referred to as a **confidence interval**. The interval informs us how certain we are about the estimate.
- **Definition:** Let $x_1, x_2, \dots, x_n$ be samples from a set of i.i.d. r.v.'s $X_1, X_2, \dots, X_n$, a parameter θ to estimate, and $0 \leq \gamma \leq 1$. If there exist sample statistics $L_n(X_1, X_2, \dots, X_n) = \mathrm{g}(X_1, X_2, \dots, X_n)$ and $U_n(X_1, X_2, \dots, X_n) = \mathrm{h}(X_1, X_2, \dots, X_n)$ such that
  $P(L_n \leq \theta \leq U_n) = \gamma$ for every value of θ.
  Then $l_n(x_1, x_2, \dots, x_n) = \mathrm{g}(x_1, x_2, \dots, x_n)$ and $u_n(x_1, x_2, \dots, x_n) = \mathrm{h}(x_1, x_2, \dots, x_n)$ form the γ-confidence interval $[l_n; u_n]$ of θ at confidence level γ for a dataset.

Confidence level?

For the case of symmetric probability distribution

Pick an interval [-*c*,*c*] such that with probability γ (e.g. 95%), the true parameter value is within this interval



Confidence interval for the mean of a normal distribution with known variance

- First consider the sample mean (our estimator) as a r.v. and transform by standardization. CLT gives that it becomes standard normal distributed $N(0,1)$ for large n.
- Our problem can then be defined as

Subtract true mean of X_n and divide by the standard deviation of the sample mean.

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = \gamma$$

$$\Rightarrow P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = \Phi(c) - \Phi(-c) = \gamma$$

Where $\Phi(c)$ is the CDF of the standard normal distribution.

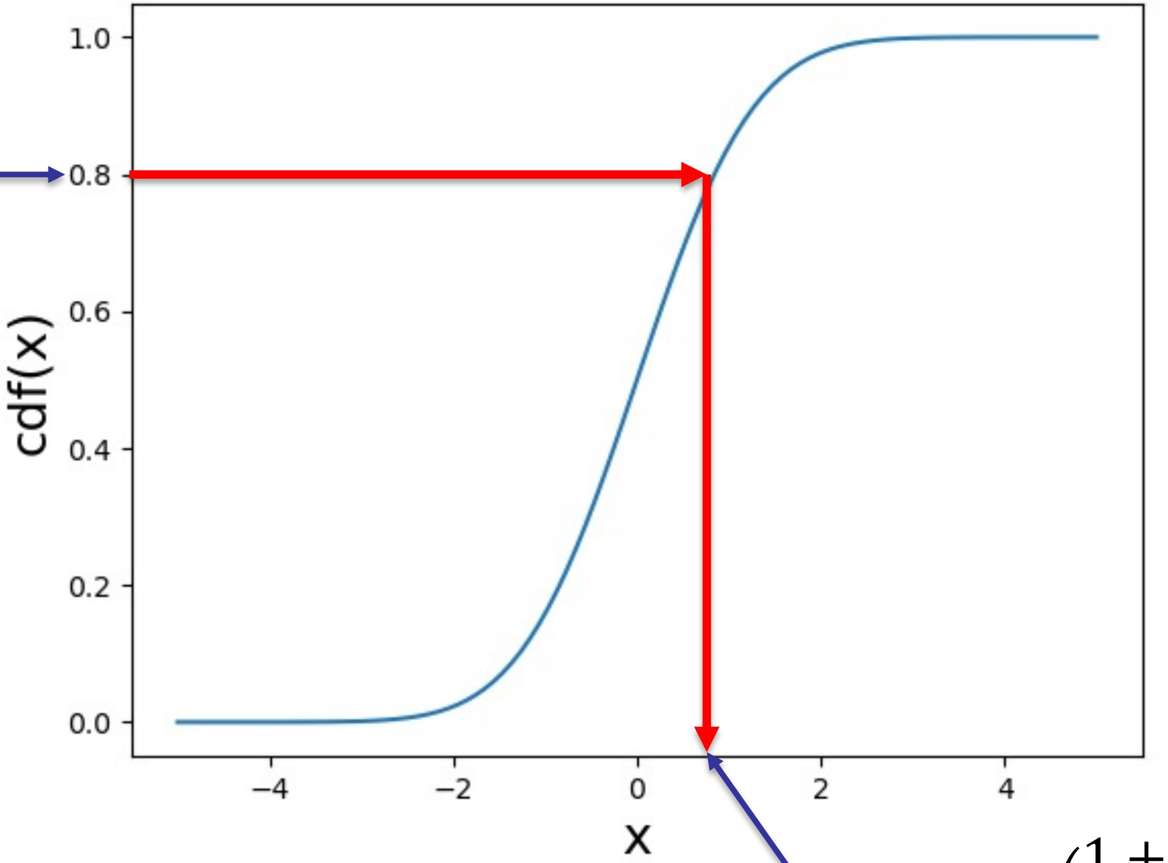
Using symmetry of $\Phi(c)$: $\Phi(-c) = 1 - \Phi(c)$

$$\Phi(c) - 1 + \Phi(c) = \gamma \Rightarrow 2\Phi(c) = 1 + \gamma \Rightarrow \Phi(c) = \frac{1+\gamma}{2} \Rightarrow$$

$$c = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

Inverse of the CDF

$$\Phi(c) = \frac{1+\gamma}{2}$$



$$c = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

Table A8 Normal Distribution

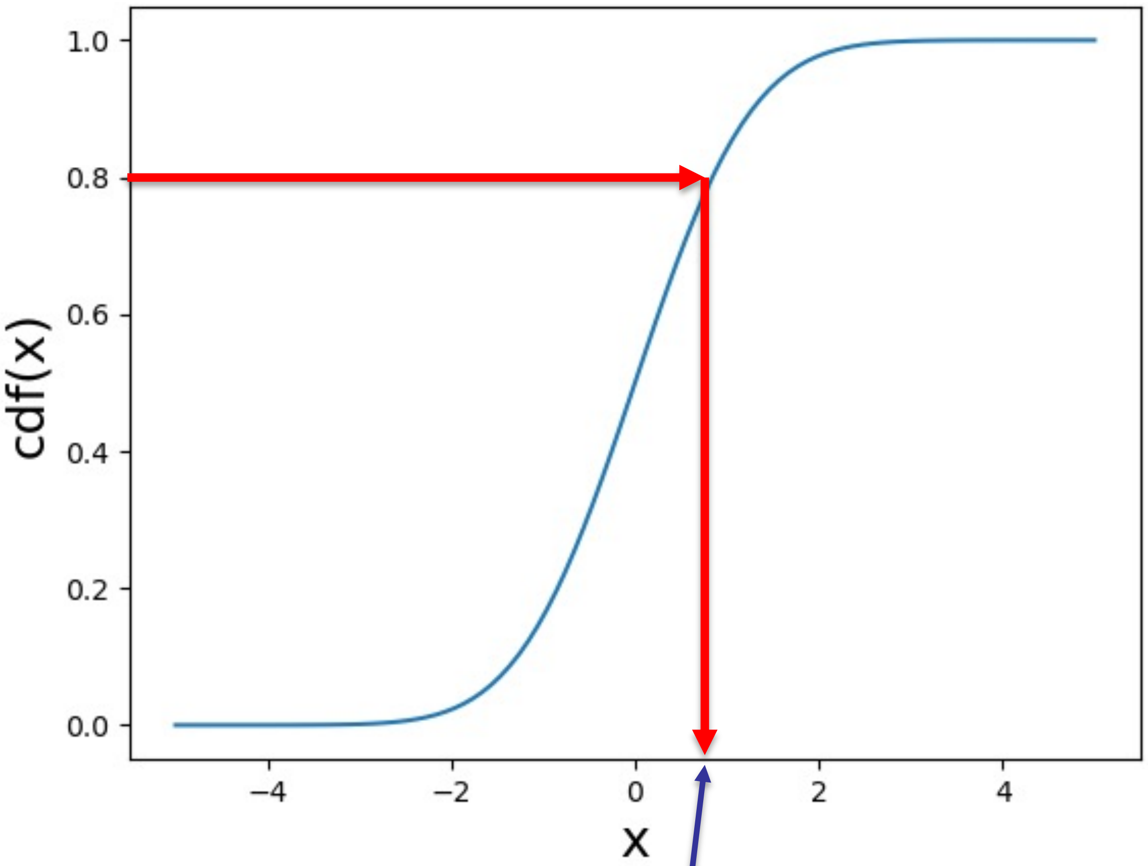
Values of z for given values of $\Phi(z)$ [see (3), Sec. 24.8] and $D(z) = \Phi(z) - \Phi(-z)$
Example: $z = 0.279$ if $\Phi(z) = 61\%$; $z = 0.860$ if $D(z) = 61\%$.

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
1	−2.326	0.013	41	−0.228	0.539	81	0.878	1.311
2	−2.054	0.025	42	−0.202	0.553	82	0.915	1.341
3	−1.881	0.038	43	−0.176	0.568	83	0.954	1.372
4	−1.751	0.050	44	−0.151	0.583	84	0.994	1.405
5	−1.645	0.063	45	−0.126	0.598	85	1.036	1.440
6	−1.555	0.075	46	−0.100	0.613	86	1.080	1.476
7	−1.476	0.088	47	−0.075	0.628	87	1.126	1.514
8	−1.405	0.100	48	−0.050	0.643	88	1.175	1.555
9	−1.341	0.113	49	−0.025	0.659	89	1.227	1.598
10	−1.282	0.126	50	0.000	0.674	90	1.282	1.645
11	−1.227	0.138	51	0.025	0.690	91	1.341	1.695
12	−1.175	0.151	52	0.050	0.706	92	1.405	1.751
13	−1.126	0.164	53	0.075	0.722	93	1.476	1.812
14	−1.080	0.176	54	0.100	0.739	94	1.555	1.881
15	−1.036	0.189	55	0.126	0.755	95	1.645	1.960

Using a table like this is old school – today we use a computer!
In python, we can compute this by

```
c = scipy.stats.norm.ppf((1+gamma)/2)
```

Percent Point Function (PPF): Inverse of the CDF



`c = scipy.stats.norm.ppf((1+gamma)/2)`

Confidence interval for the mean of a normal distribution with known variance

- The derivation was for a standardized statistics, so what is the confidence interval for the estimator $\bar{X}_n$ of μ?

$$P\left(-c \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le c\right) = \gamma$$

$$\Longrightarrow P\left(-c\frac{\sigma}{\sqrt{n}} \le \bar{X}_n - \mu \le c\frac{\sigma}{\sqrt{n}}\right) = \gamma$$

$$\Longrightarrow P\left(-\left(\bar{X}_n + c\frac{\sigma}{\sqrt{n}}\right) \le -\mu \le c\frac{\sigma}{\sqrt{n}} - \bar{X}_n\right) = \gamma$$

$$\Longrightarrow P\left(\bar{X}_n + c\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{X}_n - c\frac{\sigma}{\sqrt{n}}\right) = \gamma$$

The γ-confidence interval for the mean parameter μ of a Normal distributed sample with known variance σ^2 is $\left[\bar{X}_n - c\frac{\sigma}{\sqrt{n}} ; \bar{X}_n + c\frac{\sigma}{\sqrt{n}}\right]$.

Notice: As number of samples *n* grows the interval becomes smaller.

Steps: Confidence interval for the mean of a normal distribution with known variance.

1. Choose a confidence level γ (e.g. 95%, 99%, …)
2. Determine the corresponding critical value c:
$$c = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$
3. Compute the sample mean $\bar{x}$ of actual samples $x_1, x_2, \dots, x_n$.
4. The confidence interval for μ is
$$\left[\bar{x} - c\frac{\sigma}{\sqrt{n}}\,;\bar{x} + c\frac{\sigma}{\sqrt{n}}\right]$$
With probability γ the true mean μ will be in this interval.

Confidence interval for mean of the Normal distribution with unknown variance

- What about the more general case of mean estimator for Normal distribution with unknown variance?
- As before we consider the sample mean estimator as a r.v. and transform so it has zero mean and unit variance.
- As variance we will use the sample variance

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2$$

- Our problem can then be defined as

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \leq c\right) = \gamma$$

What is the distribution of this estimator?

- **Theorem:** Let X_1, …, X_n be i.i.d. Normal r.v.'s with mean μ and variance σ^2. Then the r.v.

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

is t-distributed with n-1 degrees of freedom (d.f). Where $\bar{X}$ is the sample mean and the sample variance is

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \bar{X} \right)^2$$

- *S* is a sum of squared Normal distributed random variables, hence *T* is t-distributed according to definition.

(Student's) t-distribution

- **Definition:** The t-distribution with *n* degrees of freedom is defined by this r.v.

$$T = \frac{Z}{\sqrt{V/n}}$$

  where $Z \sim N(0,1)$ and $V \sim \chi_n^2$ and *Z* is independent of *V*.
- Remember that the definition of the χ_n^2 distribution states that V is a sum of squared i.i.d. standard Normal distributed r.v.'s.

How to find the interval limits *c* in the case of unknown variance?

- Our problem can then be defined as

$$P\left(-c \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le c\right) = F(c) - F(-c) = \gamma$$

- Where *F*(*c*) is the CDF of the t-distribution of d.f. n-1.
- Using symmetry of the t-distribution *F*(-*c*)=1-*F*(*c*) and substition in above gives us

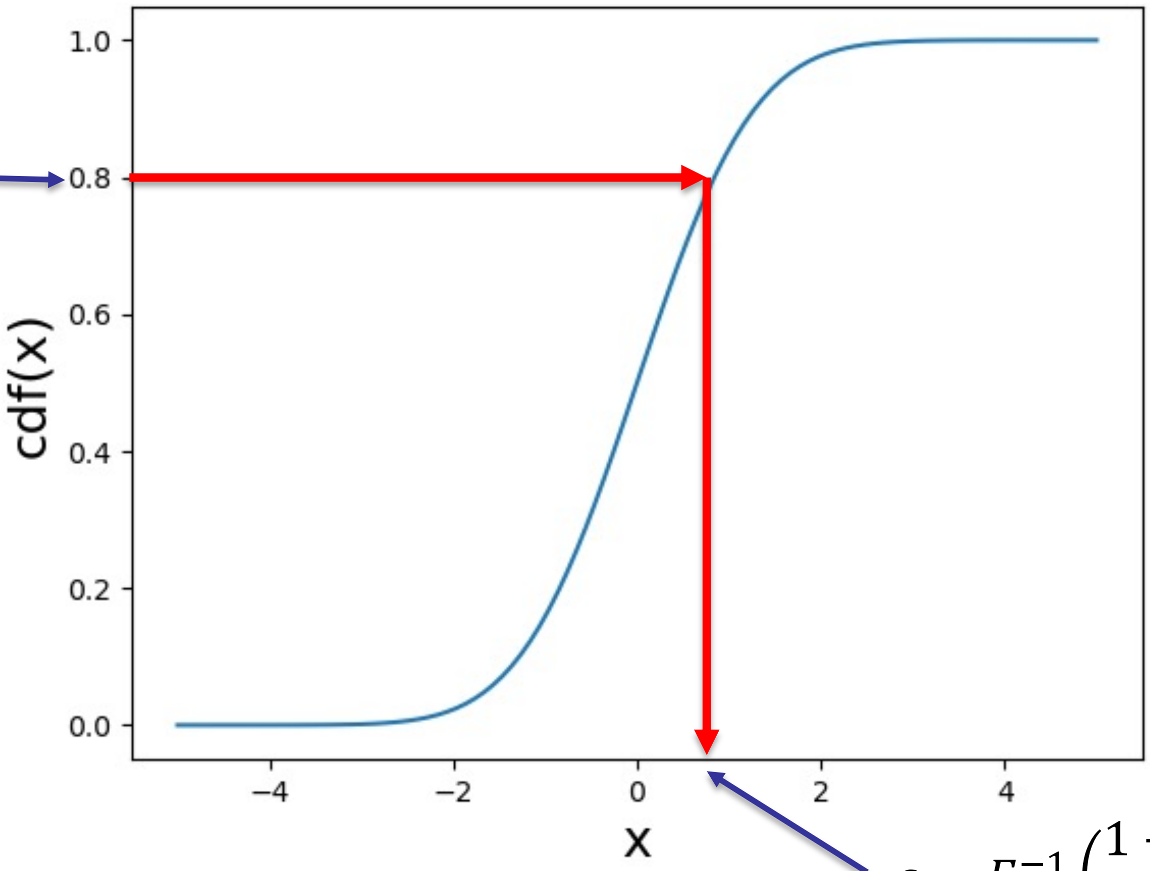
$$F(c) - F(-c) = \gamma \Rightarrow 2F(c) = 1+\gamma \Rightarrow F(c) = (1+\gamma)/2$$

- We need to compute the inverse of the CDF *F*

$$c = F^{-1}\left(\frac{1+\gamma}{2}\right)$$

Percent Point Function (PPF): Inverse lookup in the CDF

$$F(c) = \frac{1+\gamma}{2}$$



$$c = F^{-1}\left(\frac{1+\gamma}{2}\right)$$

```
c = scipy.stats.t.ppf((1+gamma)/2)
```

Confidence interval for the mean of a normal distribution with unknown variance

- We can compute the confidence interval for the estimator of μ in the same way we did for the known variance case

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \leq c\right) = \gamma$$

$$\Longrightarrow P\left(\bar{X}_n - c\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}_n + c\frac{S}{\sqrt{n}}\right) = \gamma$$

The γ-confidence interval for the mean parameter μ of a Normal distributed sample with unknown variance is $\left[\bar{X}_n - c\frac{S}{\sqrt{n}}; \bar{X}_n + c\frac{S}{\sqrt{n}}\right]$.

Notice: Again as number of samples *n* grows the interval becomes smaller.

Steps: Confidence interval for the mean of a normal distribution with unknown variance.

1. Choose a confidence level γ (e.g. 95%, 99%, …)
2. Determine the corresponding critical value *c*:

$$c = F^{-1}\left(\frac{1+\gamma}{2}\right)$$

3. Compute the sample mean $\bar{x}$ and variance s^2 of actual samples $x_1, x_2, \ldots, x_n$.
4. The confidence interval for μ is

$$\left[\bar{x} - c\frac{s}{\sqrt{n}}\,; \bar{x} + c\frac{s}{\sqrt{n}}\right]$$

   With probability γ the true mean μ will be in this interval.

Confidence intervals for parameters of other estimators

- If we have many samples, we can invoke the central limit theorem and assume that the distribution is a Normal distribution.
- This works as long as the individual r.v.'s are i.i.d. and have finite variance and our estimator is a sum of these r.v.'s (e.g. computing the sample mean).
- If so, then we can just use one of the techniques mentioned to compute confidence intervals.

Hypothesis testing

Reading material: Kreyszig, Ch. 25.4

Or

Pishro-Nik, Ch. 8.4 – 8.4.4

Hypothesis testing - intuition

- Assume we have a set of samples x_1, …, x_n from some r.v. X, and we would like to verify whether a specific hypothesis about the data is correct or not.
- **Example:** We hypothesize that the average price of food in the HCØ canteen have stayed constant since last year, where the average price was kr. 50,-. We have collected data by buying 8 meals and recording the prices. Using hypothesis testing we can evaluate whether the hypothesis holds or not.

Hypothesis testing - intuition

Pick a null and an alternative hypothesis, a significance level α, and find a critical value c based on the distribution of a test statistics (a function of the samples).

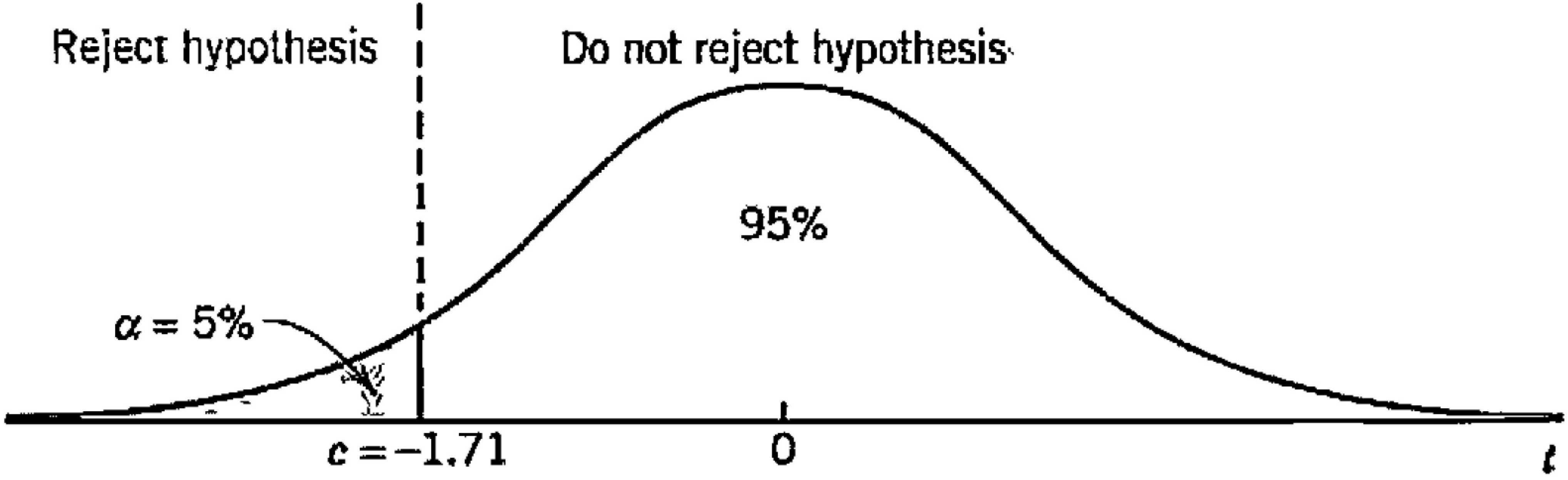


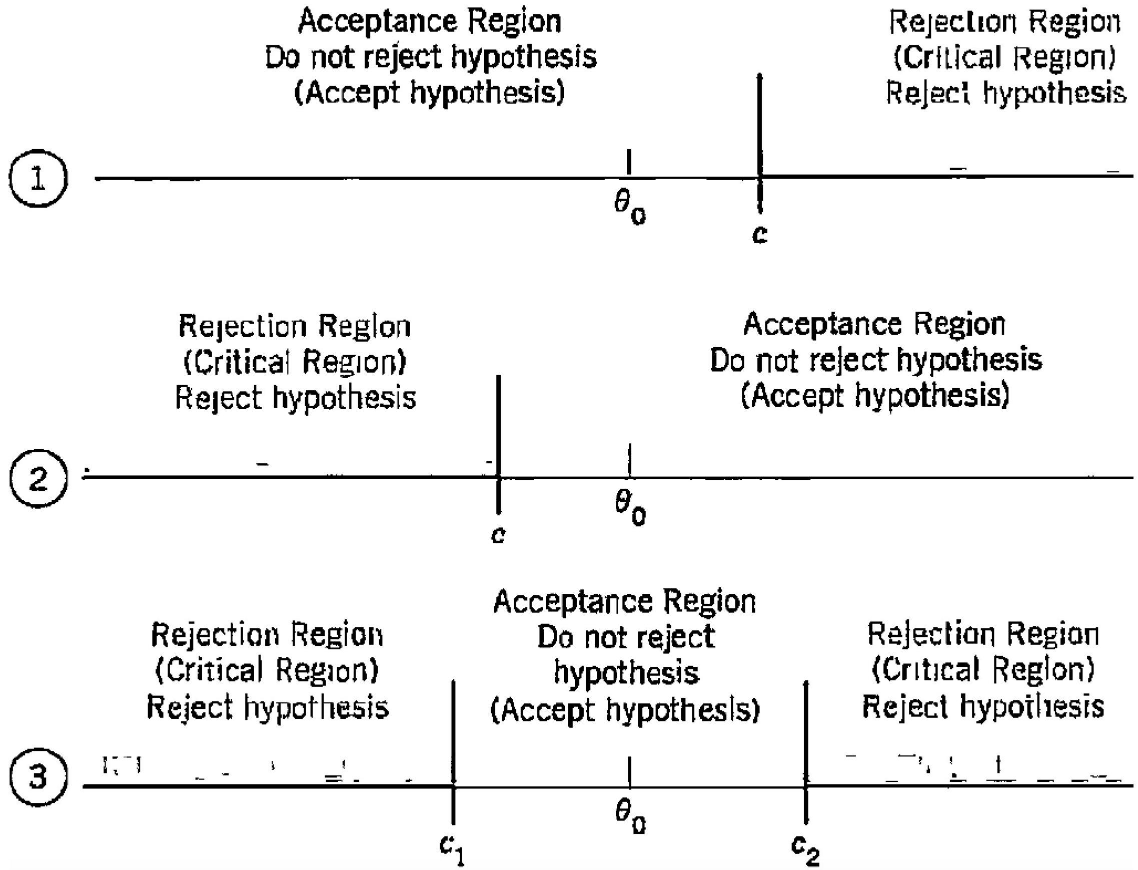
Fig. 531. t-distribution in Example 1

Types of (null) hypotheses and alternatives

Consider an unknown parameter θ and the null hypothesis is $\theta=\theta_0$. There are 3 types of alternative hypotheses

1. $\theta>\theta_0$
2. $\theta<\theta_0$
3. $\theta\neq\theta_0$

Called right-sided, left-sided, and two-sided tests.



Rejection region

- Assume we have a set of samples $x_1, \dots, x_n$ from some i.i.d. r.v. $X_1, \dots, X_n$, and a test statistics $\Theta = g(X_1, \dots, X_n)$ as well as the observed test statistics based on the samples $\theta = g(x_1, \dots, x_n)$.
- Compute the rejection regions based on the distribution of the test statistics Θ and the choice of significance level α.
- For the 3 alternative hypotheses:
  1. $P(c \leq \Theta) = 1 - P(\Theta \leq c) = \alpha \Longrightarrow P(\Theta \leq c) = 1 - \alpha$, compute the inverse of the CDF of Θ and get rejection region $\mathcal{R} = [c; \infty)$.
  2. $P(\Theta \leq c) = \alpha$, compute the inverse of the CDF of Θ and get rejection region $\mathcal{R} = (-\infty; c]$.
  3. $P(\Theta \leq c_1) = \alpha/2$ and $P(c_2 \leq \Theta) = 1 - P(\Theta \leq c_2) = \alpha/2$, and form the rejection region $\mathcal{R} = (-\infty; c_1] \cup [c_2; \infty)$.

Six steps of hypothesis testing

1. Formulate a model for the data
2. Formulate a **null hypothesis** H_0 to be tested and an **alternative hypothesis** H_A.
3. Specify a **test statistics** r.v. $\Theta = g(X_1, \ldots, X_n)$, whose distribution depends on the null and alternative hypotheses.
4. Choose a **significance level** α (e.g. 5%, 1%, 0.1%, …).
5. Compute the rejection region R based on the choice of alternative hypothesis (based on the test type).
6. Use samples $x_1, \ldots, x_n$ to compute observed value $\theta = g(x_1, \ldots, x_n)$. Reject the hypothesis, depending on if θ is in R or not (do not reject the null hypothesis).

A specific choice of test statistics: t-test

- Assume that the data is normal distributed with mean μ but unknown variance σ^2, then the relevant test statistics is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

which we, by now, know is t-distributed with n-1 d.f.'s – this is what we use to find the critical value *c*.

Example of a two-sided t-test

- **Example:** We hypothesize that the average price of food in the HCØ canteen have stayed constant since last year, where the average price was kr. 50,-. We have collected data by buying n=8 meals and recording the price. Using hypothesis testing we can evaluate whether the hypothesis holds or not.
- The observed prices are
  `x = [55, 54, 48, 75, 61, 65, 61, 49]`

Example of a two-sided t-test

- Lets assume the prices are normal distributed with mean 50 kr, but unknown variance.
- We choose the null hypothesis μ_0 = 50 kr
- The alternative is $\mu_A \neq \mu_0$, hence we have to perform a two-sided t-test.
- Lets choose the significance level to be α=5% (its not a live or death decision we are making here!)
- Our test statistics is t-distributed with d.f. n-1 = 7

$$T = \frac{\bar{X} - \mu_0}{S \big/ \sqrt{n}}$$

Example of a two-sided t-test

- The sample mean is $\bar{x} = 58.5$ kr. and sample standard deviation is $s = 8.94$ kr.
- Our test statistics is for the samples we have

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{58.5 - 50}{8.94/\sqrt{8}} = 2.69$$

- Since we are doing a two-sided t-test at significance level α=5%, we find c_1 and c_2 by inverse lookup in the t-distribution CDF at $P(T \leq c_1) = \alpha/2$ and $P(T \leq c_2) = 1 - \alpha/2$. We get $c_1 = -2.37$ and $c_2 = 2.37$.
- Since $t > c_2$ (i.e. in rejection region), we reject the hypothesis of constant price. In fact, it appears to be increasing!

Errors in the hypothesis

- A statistical test can be thought of as a decision function $d: \mathbb{R}^n \rightarrow \{H_0, H_A\}$ with $P(d = H_A) \leq \alpha$ (a bound).
- We can make two types of errors (mistakes) when doing hypothesis testing.
  - Type I error: H_0 is correct, but we choose $d = H_A$
  - Type II error: H_0 is false, but we choose $d = H_0$ anyway
- A statistical test puts a bound on the probability of making type I errors.
- Therefore, in practices always choose hypotheses $\{H_0, H_A\}$ such that the type I error becomes the ”worst error” – the one we want to avoid
- **Lunch example:** If we want to save money, it is not good if we by mistake reject constant price if it was true (type I)

Summary

- Inequalities, law of large numbers, and the central limit theorem can be used to proof various central results of probability theory and statistics.
- Inequalities are also essential in Machine Learning to proof theoretical bounds on the performance of algorithms.
- Confidence intervals provide an interval estimate of a parameter from a sample of data. The mid-point of the interval can act as point estimate and the interval as error bars on the estimate.
- Perform a statistical test of a hypothesis based on a sample of data. We looked specifically at the t-test.

Reading material

- Inequalities:
  - Blitzstein & Hwang, Ch. 10.1 (http://probabilitybook.net)
- Law of large numbers, central limit theorems, and distributions:
  - Blitzstein & Hwang, Ch. 10.2 – 10.5 (http://probabilitybook.net)
  - Pishro-Nik, Ch. 7.0 – 7.1 (https://www.probabilitycourse.com)
- Confidence intervals and hypothesis tests:
  - Kreyszig, Ch. 25.1, 25.3, 25.4
  - Pishro-Nik, Ch. 8.1 – 8.2.2, 8.3 – 8.4.4 (https://www.probabilitycourse.com)
- Supplemental reading:
  - Blitzstein & Hwang, Ch. 4.4 on indicator random variables and the fundamental bridge (needed for some proofs and in Ch. 10).