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

Plan for today

- Probability theory
  - Bounds on expectations and tail probabilities
  - Limit theorems for random variables
  - Student-t distribution and Chi-square distribution

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

How to compute expectations and probabilities?

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

Bounds using Inequalities

Cauchy-Schwarz:
A marginal bound on joint expectation

- **Theorem Cauchy-Schwarz:** For any random variables (r.v.) X and Y with finite variances,

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

where $\left|\cdot\right|$ denotes absolute value.

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

Rearranging and substitution gives

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

Hence variance is always nonnegative.

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 $g(X) = a + bX$ (g is linear) with probability 1.

- (This theorem is important within Mathematical Information Theory.)

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,

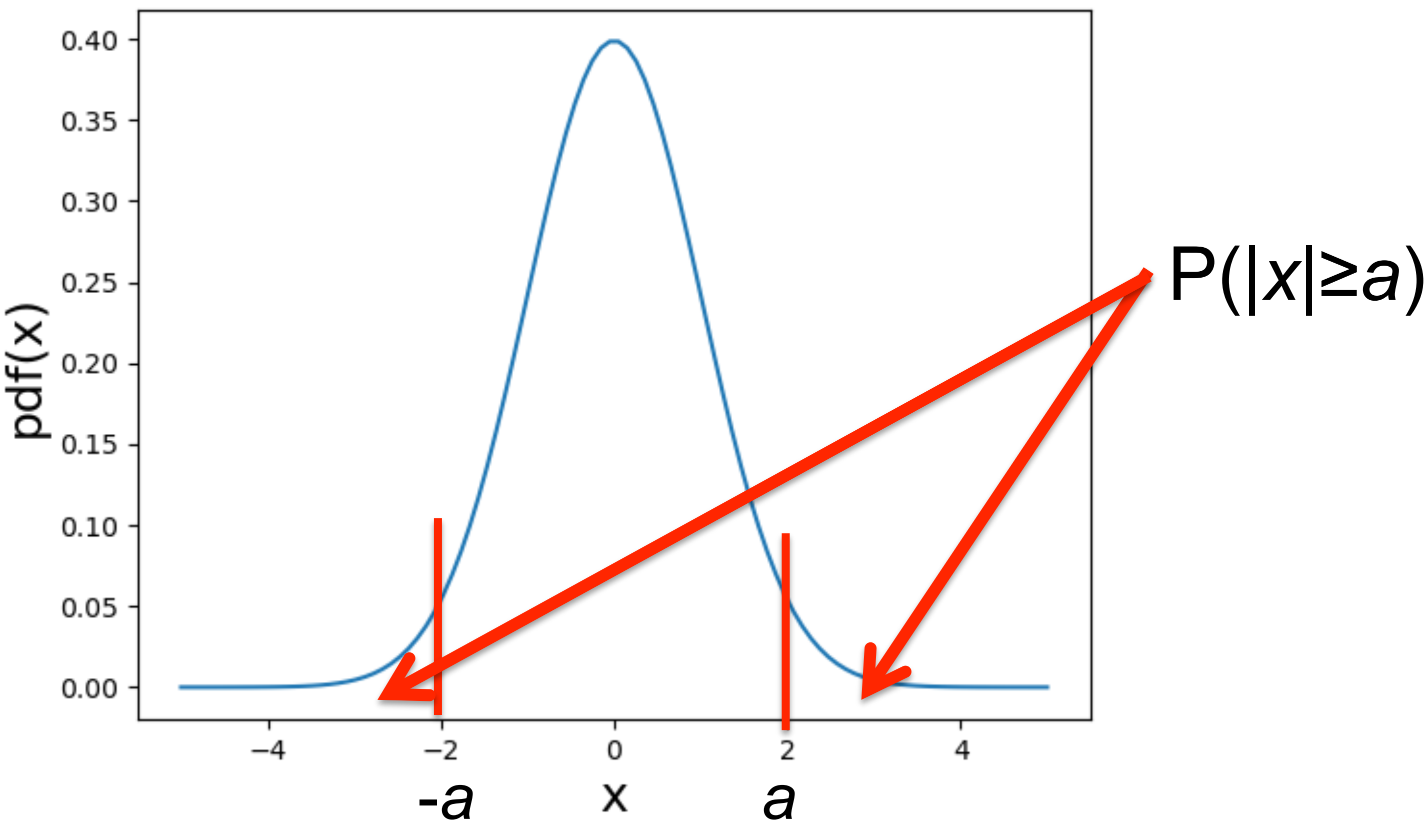
$$P\left(\left|X-\mu\right| \geq a\right) \leq \frac{\sigma^2}{a^2}$$

(A specialization of the Markov inequality)

Tail probabilities

Markov inequality for a Normal distributed r.v. *X* with N(0,σ²)

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



Convergence properties of sums of random variables

Law of Large Numbers

- 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

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

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

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

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

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

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 r.v. $\bar{X}_n$ as *n* increases?
- **Central Limit Theorem:** 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 $\bar{X}_n$

Lets look at a couple of named distributions we need now

Chi-square (χ_n^2) distribution

- **Definition:** Let $V = Z_1^2 + \ldots + Z_n^2$ where $Z_1^2, \ldots, Z_n$ 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,

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

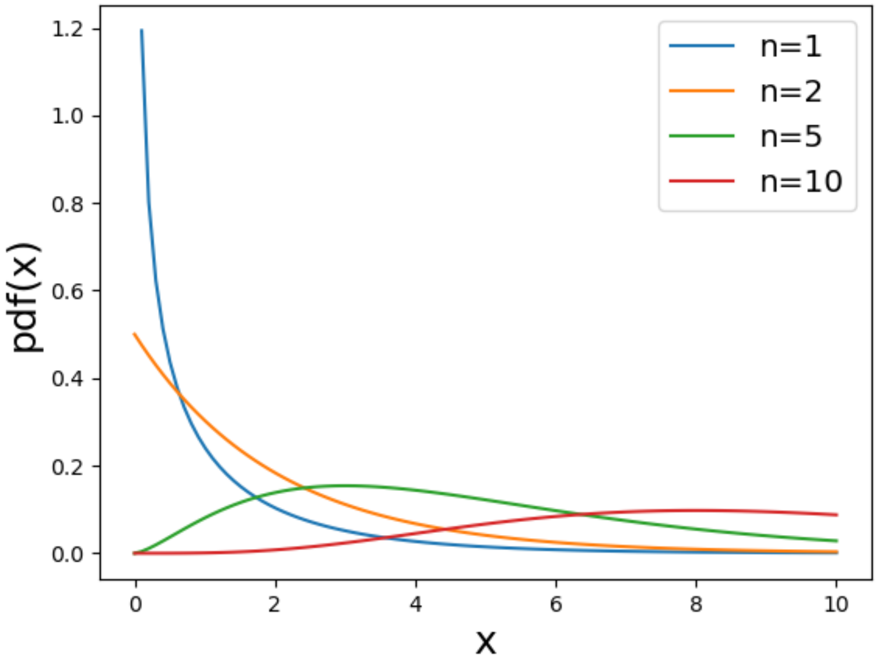
- The probability density function (PDF) is given by

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

Relates to the distribution of sample variance.

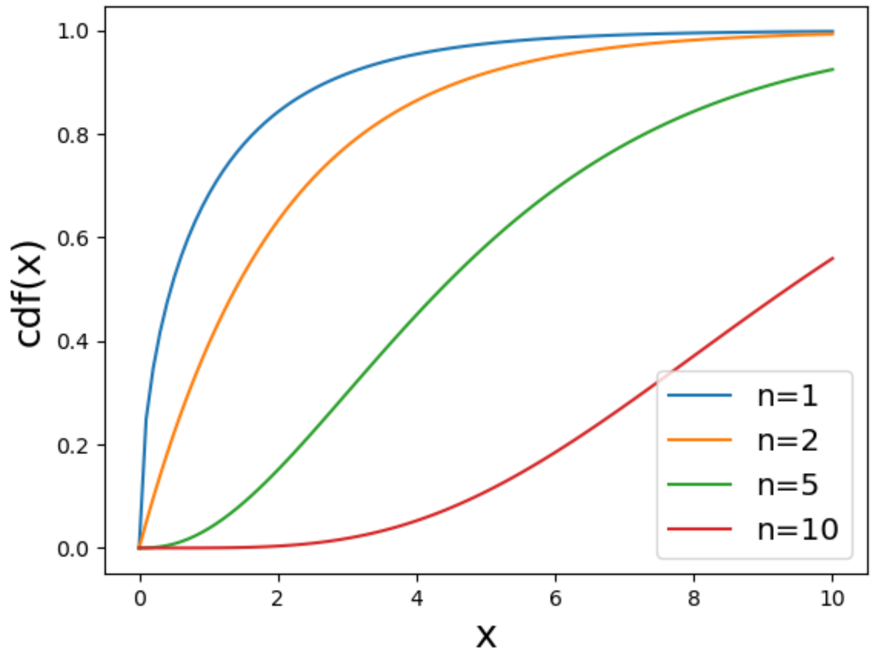
Visualizing the Chi-square (χ_n^2) distribution

PDF



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

CDF



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

(Student's) t-distribution

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

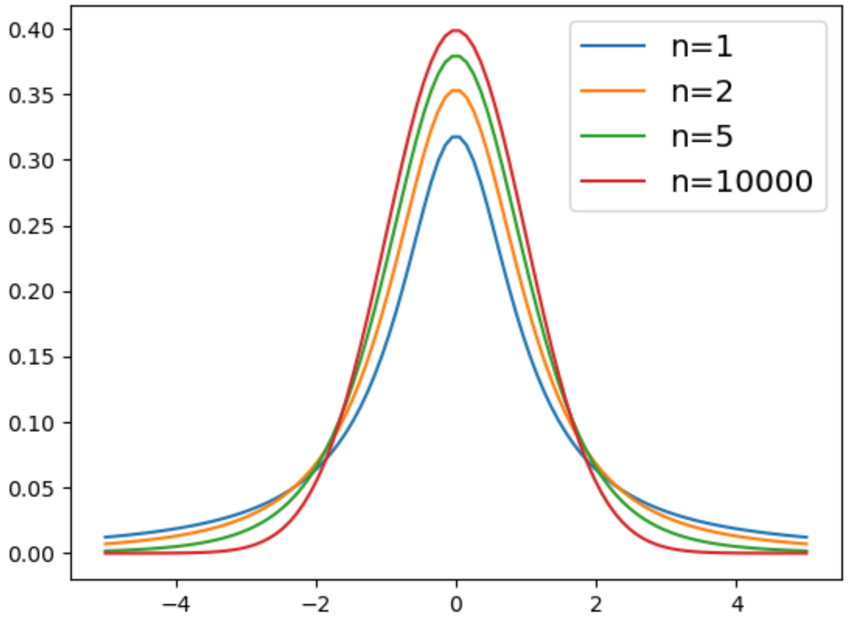
$$T = \frac{Z}{\sqrt{V/n}}$$

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

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

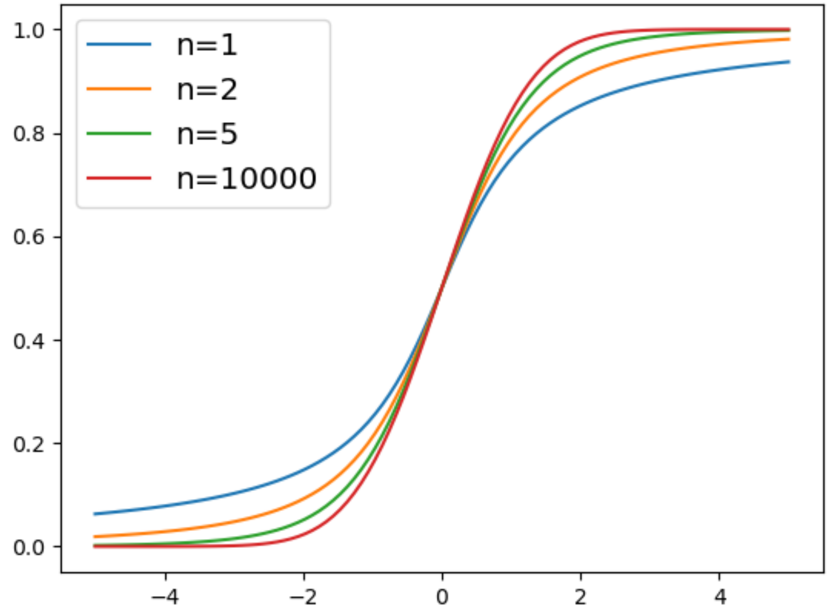
Visualizing the t-distribution

PDF



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

CDF



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

Confidence intervals and hypothesis testing

Parameter estimation

- We want to estimate parameters of a probability distribution model.
- The function computing the estimate 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 samples from a normal distributed r.v. X, then the **sample mean** is

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

Parameter estimation

- But we can also compute an interval within which the true estimate (value) lies with a chosen probability (confidence level). This is referred to as a **confidence interval**.

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

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the corresponding c: Critical value

γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

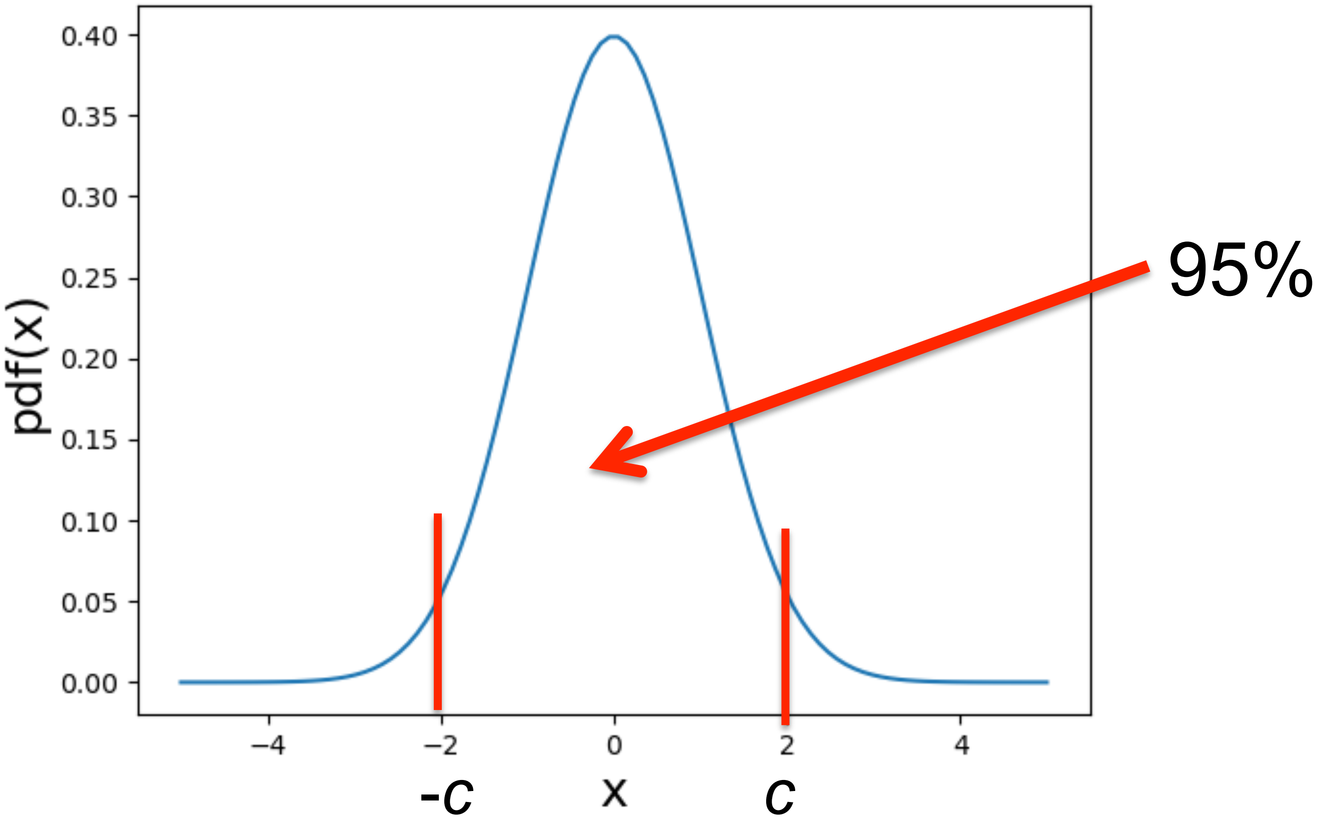
Step 3. Compute the mean $\bar{x}$ of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = c\sigma/\sqrt{n}$. The confidence interval for μ is

$$\mathrm{CONF}_\gamma \{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}.$$

Confidence level?

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



How to find the interval limits *c?*

- First consider the mean estimator as a r.v. and transform so it becomes standard normal distributed (i.e. *X*~*N*(0,1)).
- Our problem can then be defined as

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

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

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

How to find the interval limits *c*?

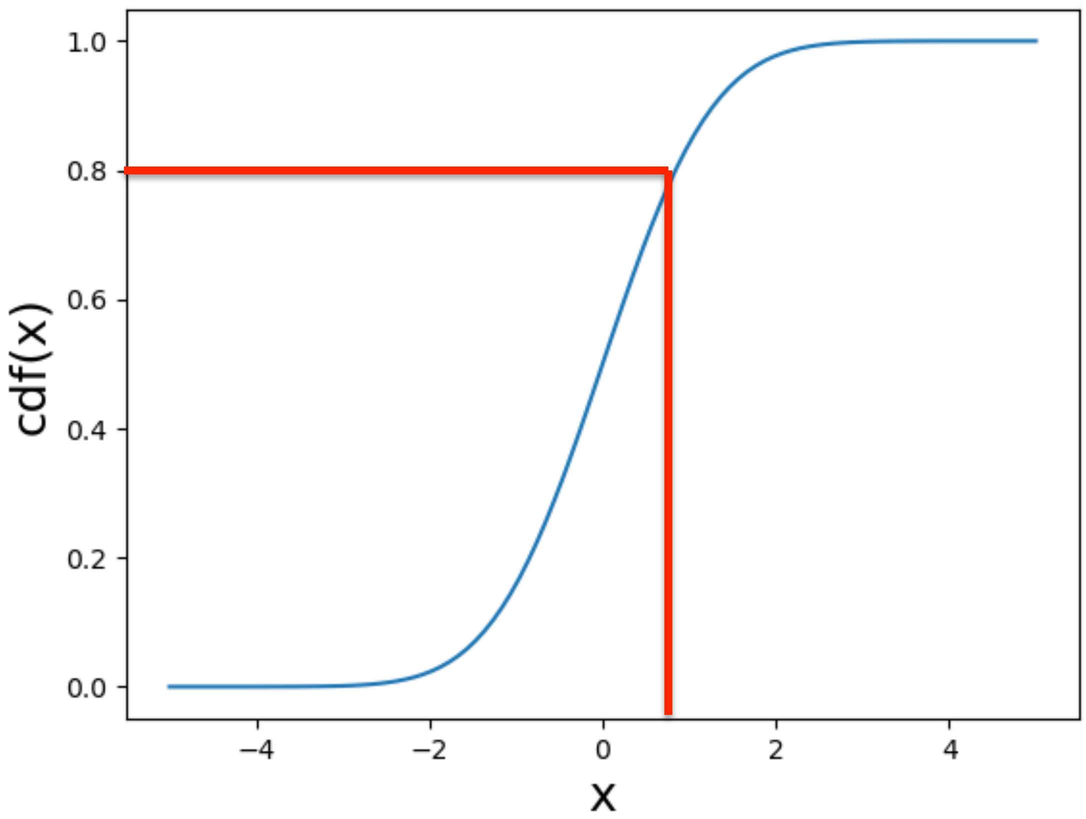
- First consider the mean estimator as a r.v. and transform so it becomes standard normal distributed (i.e. $X \sim N(0,1)$).
- Our problem can then be defined as

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

- Where $\Phi(c)$ is the CDF of the standard normal distribution.
- In python we can compute this by

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

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



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

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the corresponding c:

γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

Step 3. Compute the mean $\bar{x}$ of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = c\sigma/\sqrt{n}$. The confidence interval for μ is

$$\mathrm{CONF}_\gamma \{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}.$$

Confidence interval for mean of the Normal distribution with unknown variance

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

Whats this?

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

from the table of the t-distribution with $n - 1$ degrees of freedom (Table A9 in App. 5; or use a CAS; n = sample size).

Step 3. Compute the mean $\bar{x}$ and the variance s^2 of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = cs/\sqrt{n}$. The confidence interval is

$$\mathrm{CONF}_{\gamma}\,\{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}.$$

Why do we need the t-distribution for the critical value?

- **Theorem:** Let $X_1, \ldots, X_n$ 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

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

How to find the interval limits *c?*

- As before we consider the mean estimator as a r.v. and transform so it zero mean and unit variance.
- Our problem can then be defined as

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

- 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

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

- In python we can compute this by

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

Confidence interval for mean of the Normal distribution with unknown variance

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

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

from the table of the t-distribution with $n - 1$ degrees of freedom (Table A9 in App. 5; or use a CAS; n = sample size).

Step 3. Compute the mean $\bar{x}$ and the variance s^2 of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = cs/\sqrt{n}$. The confidence interval is

$$\text{CONF}_{\gamma}\,\{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}.$$

Confidence intervals for parameters of other distributions

- No problem! We just need to invoke the central limit theorem and use more samples.
- 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.

The Central Limit Theorem (Recap)

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

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

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

Hypothesis testing

Hypothesis testing - intuition

- Assume we have a set of samples $x_1, \ldots, x_n$ from some r.v. X, and we would like to verify whether a specific assumption about the data is correct or not.
- **Example:** We hypothesize that the average price of food in the Biocenter 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 price. Using hypothesis testing we can evaluate whether the hypothesis holds or not.

Hypothesis testing - intuition

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

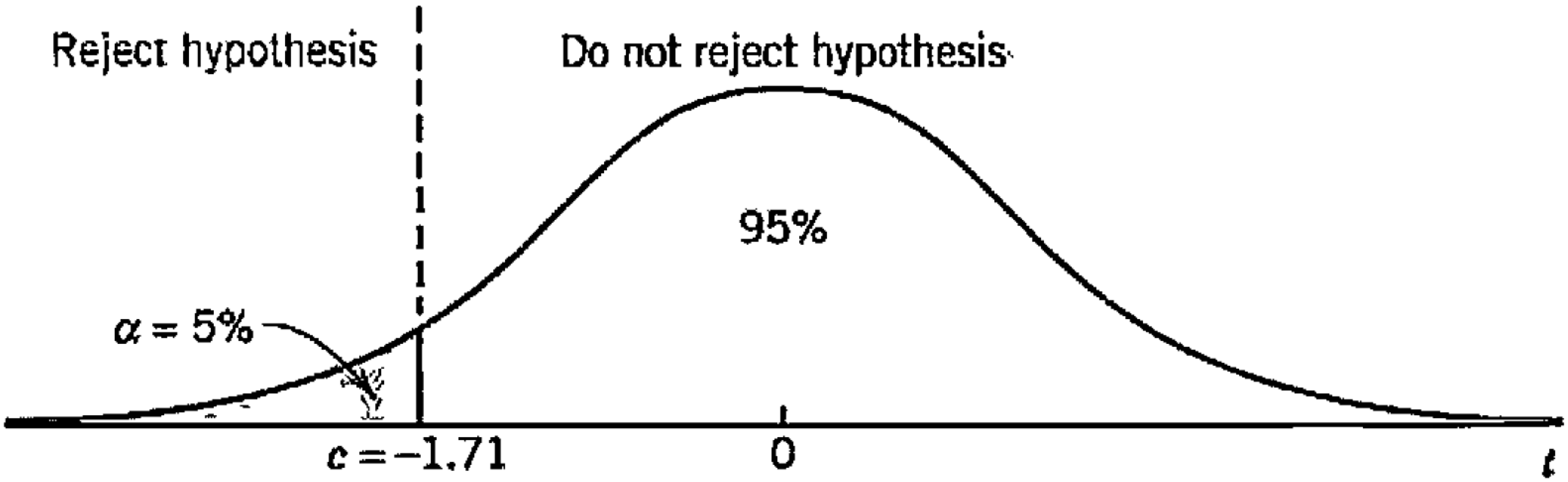
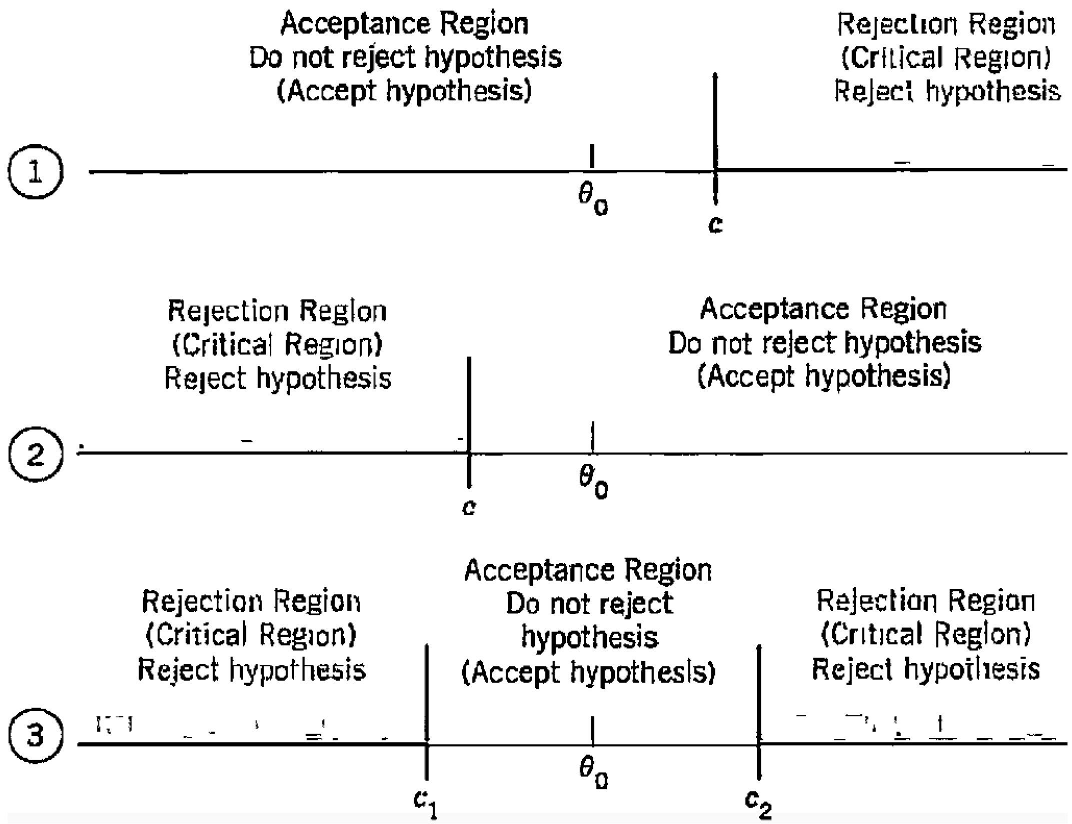


Fig. 531. *t*-distribution in Example 1

Types of (null) hypotheses and alternatives

- Consider an unknown parameter θ and the null hypothesis $\theta=\theta_0$. There are 3 types of alternatives
  $\theta>\theta_0$
  $\theta<\theta_0$
  $\theta\neq\theta_0$
  Called right-sided, left-sided, and two-sided tests.



6 steps of hypothesis testing

1. Formulate a **hypothesis** $\theta=\theta_0$ to be tested (also called the **null hypothesis**)
2. Formulate an **alternative** $\theta=\theta_1$.
3. Choose **significant level** α (e.g. 5%, 1%, 0.1%, …).
4. Use a r.v. $\Theta=g(X_1, \ldots, X_n)$ as test statistics, whose distribution depends on the hypothesis and alternative. Compute a critical value c based on this distribution by $P(\Theta \leq c)=\alpha$.
5. Use samples $x_1, \ldots, x_n$ to compute observed value $\theta=g(x_1, \ldots, x_n)$.
6. Accept or reject the hypothesis, depending on the size of θ relative to c based on the choice of test type.

T-test (a specific choice of test statistics)

- Assume that the data is normal distributed with known 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 Biocenter 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 $\mu_1 \neq \mu_0$, 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

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

Example of a two-sided t-test

- The sample mean is $\bar{x} = 58.5$ and sample standard deviation is $s = 8.94$
- 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 $t > c_2$, we reject the hypothesis of constant price. In fact, it is increasing!

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, law of large numbers, central limit theorems, and distributions:
  - Blitzstein & Hwang, Ch. 10.1 – 10.5
- Confidence intervals and hypothesis tests:
  - Kreyszig, Ch. 25.1, 25.3, 25.4
- Supplemental reading:
  - Blitzstein & Hwang, Ch. 4.4 on indicator random variables and the fundamental bridge (needed for some proofs and examples in Ch. 10).