**BI-SPOL-26 Základy statistické indukce, náhodný výběr, bodové odhady pro střední hodnotu a rozptyl, intervalové odhady pro střední hodnotu, testování statistických hypotéz o střední hodnotě**

BI-PST

## Úvod do statistiky

* Na základě skutečných výsledků (naměřených dat) vybíráme vhodný model, odhadujeme hodnoty jeho parametrů, testujeme hypotézy o těchto parametrech a ověřujeme shodu modelu se skutečností.
* Postupujeme do jisté míry opačně než v teorii pravděpodobnosti – tam jsme vytvořili matematický model a poté ho používali k predikci možných výsledků (tj. určování pravděpodobnosti jevů, rozdělení a stření hodnoty náhodných veličin)
* Máme například urnu s neznámým počtem červených a modrých kuliček. Na základě náhodného výběru budeme odhadovat podíl červených a modrých kuliček, testovat, zda je v urně 50% modrých kuliček nebo více atd. V teorii pravděpodobnosti bychom věděli, že máme r červených a b modrých a počítali bychom, s jakou pravděpodobností vytáhneme modrou, apod.

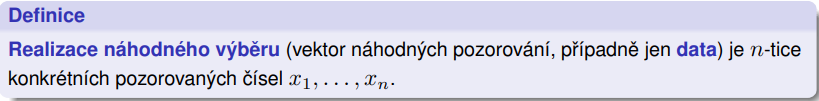
## Náhodný výběr

* Nevíme, jak data „vypadají“ (než je naměříme), proto je budeme modelovat pomocí náhodných veličin
* Když si dnes naplánuji, že budu zítra provádět experiment, tak dnes je to náhodný výběr, ale zítra se to stává realizací náhodného výběru – naměřená data



Obsah obrázku text

Popis byl vytvořen automaticky



## Základy statistické indukce

* metoda, která dovoluje odhadnou chování celku na základě jeho vybraných částí (náhodný výběr) – z výběrových dat usuzujeme obecnější skutečnosti
* **máme náhodný výběr z neznámého rozdělení. Na základě naměřených dat (realizace náhodného výběru) chceme o tomto rozdělení zjistit co nejvíce**
* typické kroky statistického zkoumání:
* **odhad tvaru rozdělení**
* **odhad parametrů rozdělení** (normální má dva – střední hodnotu a rozptyl)
* **ověření správnosti modelu – testování hypotéz**



* např. binomické rozdělení

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

### Odhad tvaru rozdělení

**Volba modelu**

Obsah obrázku text, interiér, snímek obrazovky

Popis byl vytvořen automaticky

Obsah obrázku text

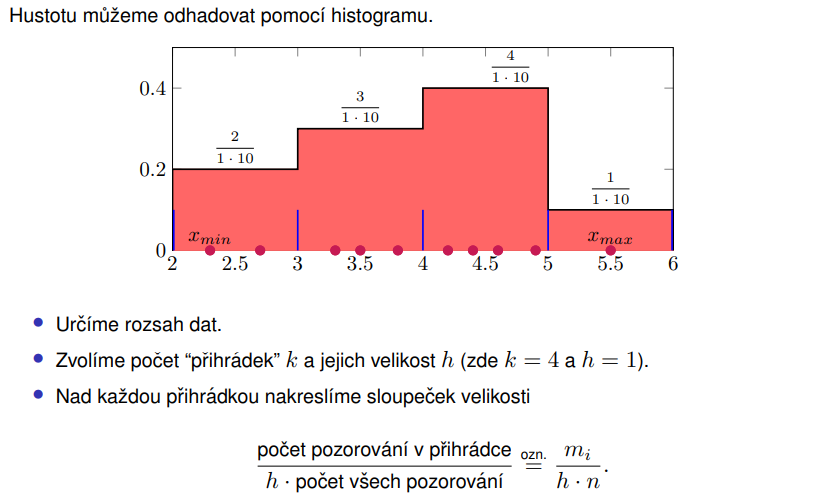
Popis byl vytvořen automaticky

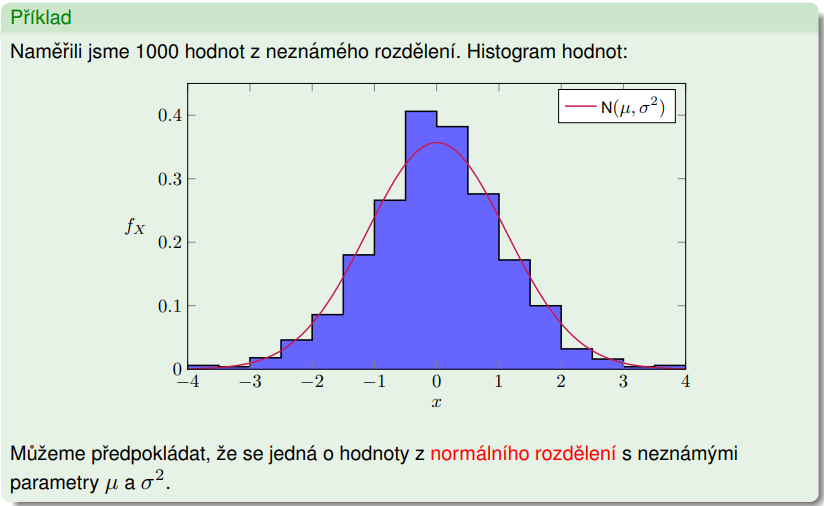
* **Příklady možných modelů**

Obsah obrázku text

Popis byl vytvořen automaticky

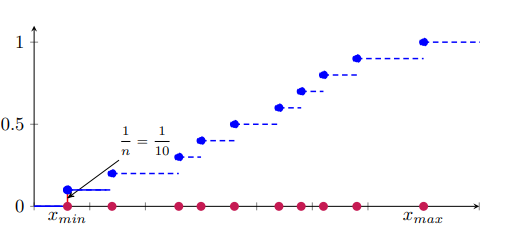
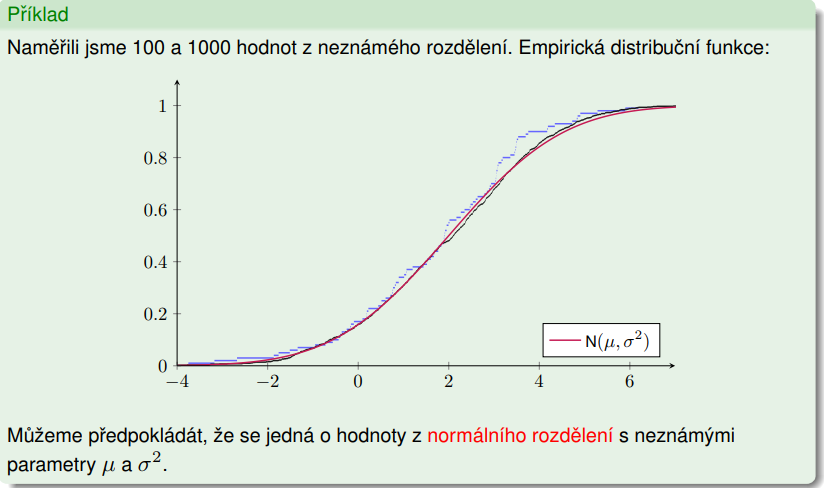
**Histogram**





**Empirická distribuční funkce**

* Pomocí empirické distribuční funkce odhadujeme distribuční funkci
* Pravděpodobnost, že zkoumaná veličina bude menší nebo rovná *x*, můžeme odhadnout jako podíl pozorovaných dat, která jsou menší nebo rovná *x*



**Příklad**

Strana 113 – 115

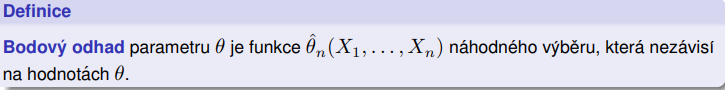
**Kvantily**

Obsah obrázku text

Popis byl vytvořen automaticky

## Bodové odhady

* aproximace hodnot parametru θ z naměřených dat



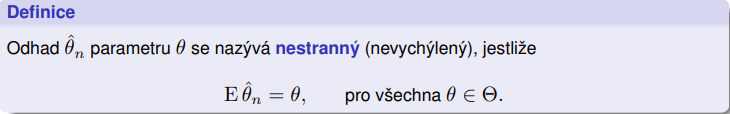
### Vlastnosti bodových odhadů

Obsah obrázku text

Popis byl vytvořen automaticky

* U odhadu požadujeme, aby byl **nestranný** a **konzistentní**

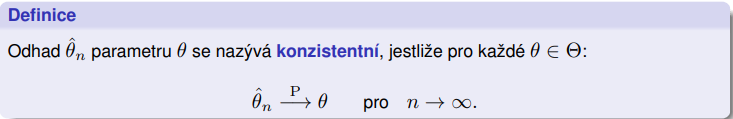
**Nestrannost**



* Nestrannost znamená, že odhad není zatížen systematickou chybou, tedy že odhadovanou hodnotu systematicky nepřestřelujeme ani nepodhodnocujeme
* Střední hodnota funkce je rovna parametru

**Konzistence**

* Volbou velkého počtu náhodných veličin (*n do nekonečna)* lze učinit chybu dostatečně malou
* Odhad konverguje v pravděpodobnosti k parametru θ (k tomu, co chci odhadnout)



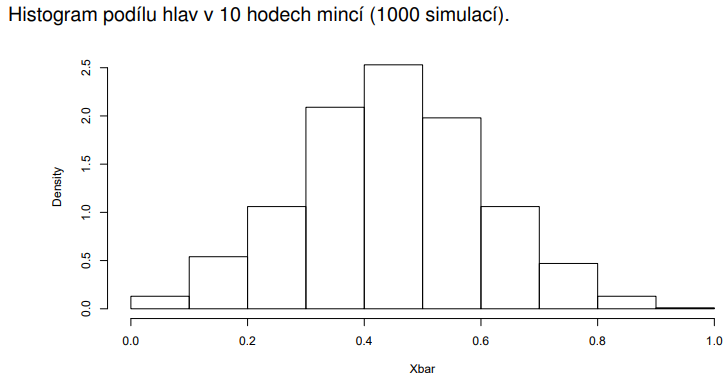
Obsah obrázku text

Popis byl vytvořen automaticky

* Pokud je stření hodnota čtverce odhadu konečná a střední hodnota odhadu konverguje k θ (odhad je asymptoticky nestranný) a rozptyl odhadu konverguje k 0, pak je odhad konzistentní
* (bokem), (není zas tak důležité)

### Bodový odhad střední hodnoty

* Bodový odhad střední hodnoty je **výběrový průměr**
* Klasický aritmetický průměr
* Nestrannost platí
* Konzistence platí ze slabého zákona velkých čísel



### Bodový odhad rozptylu

* Bodový odhad rozptylu je **výběrový rozptyl**
* Součet rozptylů dělený počtem *n - 1* (aby byla zajištěna nestrannost), odchylka (odmocnina z rozptylu) je vlastně „vzdálenost“ mezi hodnotou a střední hodnotou
* Nestrannost platí
* Konzistence platí ze zákona velkých čísel

### Kvalita nestranného odhadu

Obsah obrázku text

Popis byl vytvořen automaticky

### Metoda momentů

* Metoda pro hledání odhadů (viz. papíry)

Obsah obrázku text

Popis byl vytvořen automaticky

### Metoda maximální věrohodnosti

* Metoda pro hledání odhadů (viz. papíry)

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Existují další bodové odhady (výběrová směrodatná odchylka, výběrová kovariance, *k*-tý výběrový moment, výběrový korelační koeficient…)

### Rekapitulace bodových odhadů

Obsah obrázku text

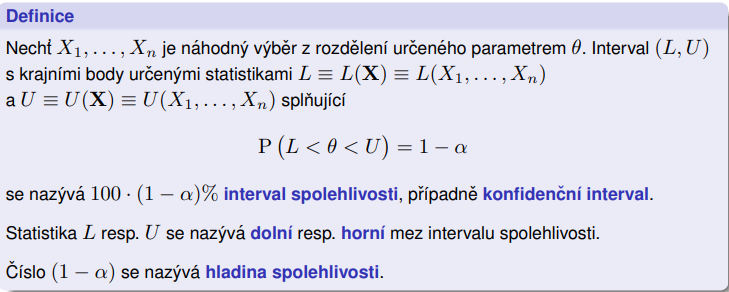
Popis byl vytvořen automaticky

## Intervalové odhady

* Místo bodového odhadu parametru θ nás může zajímat interval, ve kterém leží skutečná hodnota parametru s danou vysokou pravděpodobností 1 – α

### Interval spolehlivosti

* Sestavíme meze L (lower bound) a U (upper bound) jako nějaké funkce náhodného výběru, takové aby pravděpodobnost, že ten parametr, který chci odhadnout byl mezi nimi, byla 1 – α, kde α je nějaké dané malé číslo
* L a U jsou funkce těch jednotlivých X ( L(X1, X2, X3, …, Xn) ). Píšu L a U, ale myslím tím funkce



* jinak řečeno: máme jistotu 100·(1 − α)%, že parametr θ se v tomto intervalu skutečně nachází (lze sestavit i pouze dolní/horní interval spolehlivosti)

Obsah obrázku text

Popis byl vytvořen automaticky

**Intuitivně**: Místo toho, abychom hledali konkrétní hodnotu parametru, tak pouze hledáme interval (*L, U*), ve kterém bude ležet. Přičemž pravděpodobnost má být např. oboustranná 90%. Tzn. budeme mít osu x rozdělenou jako 5% - L – 90% - U – 5%. Tzn. výsledkem může být (25.111, 25.222) – někde zde leží parametr s pravděpodobností 90%.

**Jednostranné intervaly spolehlivosti**

**Obsah obrázku text

Popis byl vytvořen automaticky**

**Postup sestavení intervalů spolehlivosti**

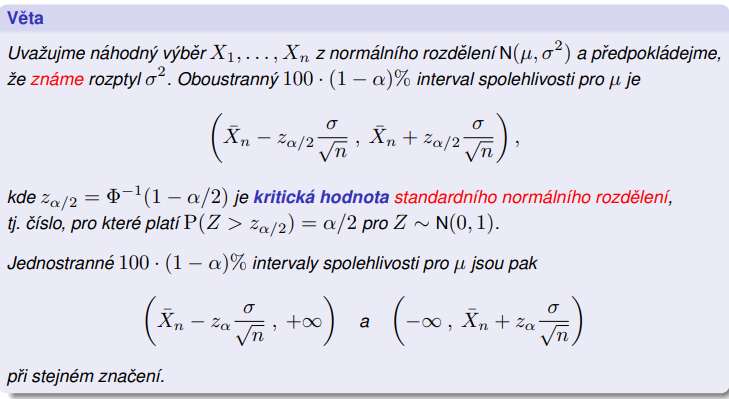
* Obecný postup:
  + **Najděme statistiku H(θ),** která: závisí na odhadovaném parametru θ, závisí na náhodném výběru X1, .., Xn a má známé rozdělení
  + Nalezneme takové meze hL a hU, pro které P(hL < H(θ) < hU) = 1 – α
  + Úpravou nerovností osamostatníme θ a získáme P(L < θ , U) = 1 - α

### Intervalové odhady pro střední hodnotu

* potřebujeme intervalový odhad (L, U), ve kterém leží střední hodnota. Takový interval se nazývá interval spolehlivosti
* záleží známe rozptyl nebo ne

**Známý rozptyl**

* známe rozptyl a chceme zjistit (odhadnout) střední hodnotu
* vezmeme výběrový průměr a přičítáme a odečítáme
* (kritická hodnota v bodě α/2 je kvantil k bodu 1 – α/2) – číslo, které separuje horní α/2 část rozdělení
* Při známém rozptylu bereme kritické hodnoty z normálního rozdělení



* gaussovo rozdělení, standardizovali jsme výběrový průměr
* lze použít libovolné rozdělení díky CLV (pro velké *n* a pouze přibližně)
* kritická hodnota standardního normálního rozdělení je v tabulkách
* α určuje konfidenční interval, čím vyšší α, tím užší interval spolehlivosti, protože je méně spolehlivý (např. α = 0,2 ⇒ 80% interval spolehlivosti, může být užší)

### 

* kritické hodnoty nám říkají, v jakých mezích se bude pohybovat standardizovaný průměr

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

* Příklad pro hodnoty jiné než normální rozdělení

Obsah obrázku text

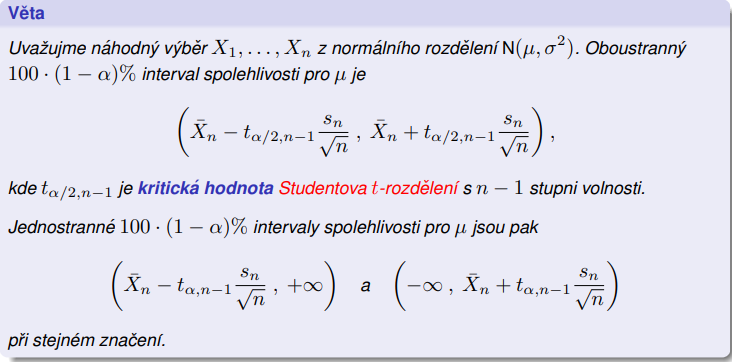
Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

**Neznámý rozptyl**

* Často rozptyl neznáme a máme k dispozici jen pozorovaná data
* rozptyl nahradíme (odhadneme) odmocninou z výběrového rozptylu
* při neznámém rozptylu bereme kritické hodnoty ze studentova rozdělení



* studentovo rozdělení
* opět lze použít na libovolné rozdělení díky CLV (pro velké *n* a pouze přibližně)
* **přibližná** spolehlivost intervalu je pak 1 – α
* intervaly jsou širší, než pro známý rozptyl (protože kalkulujeme s jeho odhadem)
* dostatečný rozsah výběru obvykle znamená *n* = 30 nebo *n* = 50. V případě rozdělení výrazně odlišných od normálního (více vrcholů hustoty, šikmost) musí často být *n* ještě vyšší

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

## Testování statistických hypotéz o střední hodnotě

* často potřebujeme ověřit tvrzení o zkoumaném rozdělení dat, ale máme k dispozici jen náhodný výběr (např. vadné produkty při přijetí)

### Hypotézy

**Obsah obrázku text

Popis byl vytvořen automaticky**

**Obsah obrázku text

Popis byl vytvořen automaticky**

**Obsah obrázku text

Popis byl vytvořen automaticky**

### Chyby testování

Obsah obrázku text

Popis byl vytvořen automaticky

### Možné výsledky:

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

### Parametrické testy a intervaly spolehlivosti

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

**Parametrické testy – ilustrace**

Obsah obrázku text

Popis byl vytvořen automaticky

### Postup

* Zvolíme hladinu významnosti α
* Napozorujeme realizace náhodného výběru
* Zkonstruujeme konfidenční interval o spolehlivost 1 – α, odpovídající alternativní hypotéze
* Zamítneme hypotézu H0 ve prospěch HA na hladině významnosti α, pokud testovaná hodnota neleží v tomto intervalu spolehlivosti

### Testy o parametrech normálního rozdělení

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

**Příklad**

Obsah obrázku text

Popis byl vytvořen automaticky

### Testování hypotéz založené na CLV

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

### Testování pomocí testových statistik

* Od tohoto níž se už učit jen okrajově – že něco takové existuje a základní princip
* Tato metoda je rozšiřitelná na testování složitějších úloh

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku stůl

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

**p-hodnota**

* Nejnižší hladina významnosti, na které bychom zamítli hypotézu

Obsah obrázku text

Popis byl vytvořen automaticky

**Párový test**

Obsah obrázku text, osoba, snímek obrazovky, dokument

Popis byl vytvořen automaticky

Obsah obrázku stůl

Popis byl vytvořen automaticky

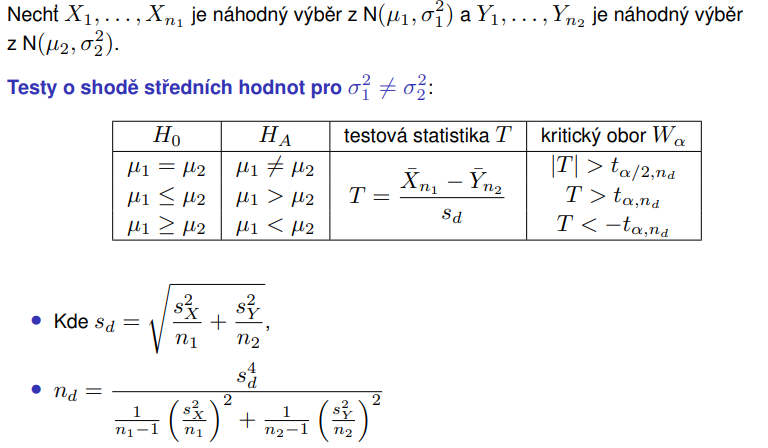
**Dvouvýběrový t-test**

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky



Obsah obrázku text

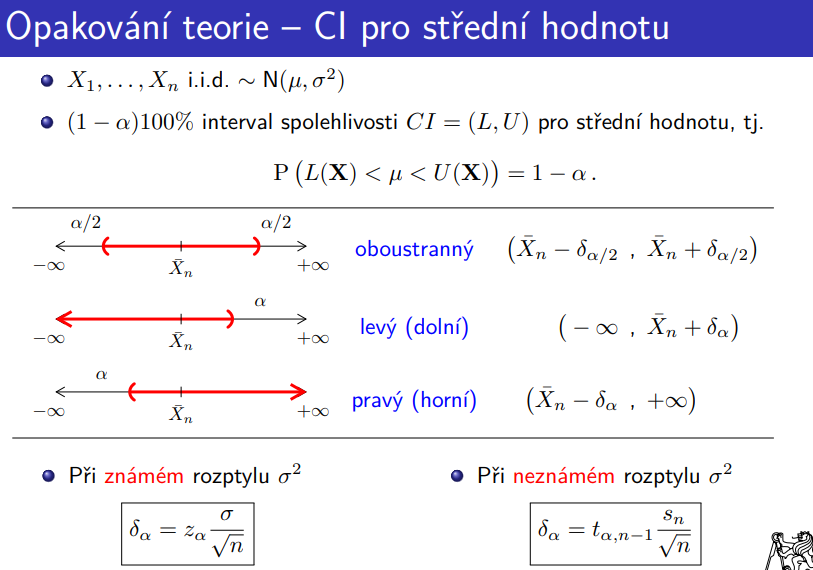
Popis byl vytvořen automaticky

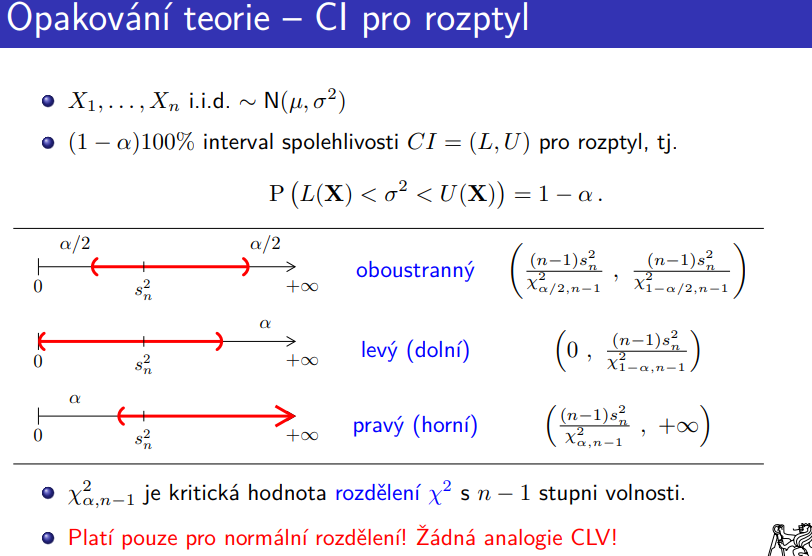
## Rekapitulace

Obsah obrázku text

Popis byl vytvořen automaticky

## Otázky a odpovědi

1. Co je vlastně kritická hodnota normálního rozdělení?
2. Uvést vzorečky pro normální, studentovo a chí-kvadrát  
   normální a studentovo – pro střední hodnotu u intervalu  
   

chí-kvadrát – pro rozptyl u intervalu  


1. K čemu je chí-kvadrát?  
   chí-kvadrát je rozdělení, jejích kritické hodnoty se používají pro intervalové odhady rozptylu
2. Co je studentovo rozdělení a chí-kvadrát?  
   Jsou to teoretická rozdělení  
   studentovo se používá u intervalových odhadů střední hodnoty  
   chí-kvadrát se používá u intervalových odhadů rozptylu
3. Co je kritický obor?  
   To jsou okrajové (odlehlé) části grafu – intervaly (- nekonečno, ), (, + nekonečno)
4. Princip testování

**Poznámky:**

Normální rozdělení je limitou rozdělení studentova