Základní statistické modely

Statistika

- Matematická statistika se zabývá interpretací získaných náhodných dat.
- Snažíme se přiřadit statistickému souboru vhodnou distribuční funkci a najít základní číselné parametry popisující pravděpodobnostní chování náhodného procesu generujícího náhodná data.
- **Základní soubor** mi představuje všechny možné hodnoty, kterých může sledovaná náhodná veličina nabývat – konečný nebo nekonečný. Zajímají nás jeho vlastnosti.
- **Výběrový (statistický) soubor** mi představuje podmnožinu konečného množství *n* prvků ze základního souboru. Typicky vznikne na základě experimentu (měření) náhodné veličiny, kterou mi generuje nějaký náhodný proces. Výběrové statistiky popisují jen **výběrový soubor!!!**
- Vhodný **statistický model** výběrového souboru mi pak umožní usuzovat na vlastnosti základního souboru. Tedy výběrový soubor musí být „dobrým“ reprezentantem základního souboru.
- Prvky výběrového souboru musí být získány náhodně a nesmí tedy na sobě vzájemně záviset.

Výběrový soubor

- Příkladem statistického souboru je měření rychlosti světla *c* Michelsonem. V tabulce jsou změřené hodnoty *c* plus 299 000 km/s.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

Náhodný výběr

- Jednotlivá měření jsou realizace náhodné proměnné X_i, i = 1, 2, 3, …, n mající všechny stejnou pravděpodobnostní distribucí a zároveň jsou nezávislé.
- Takovou kolekci náhodných proměnných nazýváme jako **náhodný výběr**.

> RANDOM SAMPLE. A *random sample* is a collection of random variables $X_1, X_2, \ldots, X_n$, that have the same probability distribution and are mutually independent.

- Pokud F je distribuční funkce náhodné proměnné X_i, pak mluvíme o náhodném výběru z F.
- V mnoha situacích jsme schopni stanovit F (dopředu ji známe), z níž generujeme náhodný výběr (např. Poissonovské procesy).
- Na druhou stranu existují náhodné procesy, jejichž pravděpodobnostní distribuci neznáme (doba trvání erupce gejzíru – viz přednáška 5).
- Každopádně je jisté, že jsme získali soubor dat – **statistický (výběrový) soubor** – jako opakování stejného měření za stejných experimentálních podmínek.

Statistický model

- Základní **statistický model** pro takový statistický soubor je:
  - považovat měření jako **náhodný výběr**
  - a interpretovat **statistický soubor** jako realizaci náhodného výběru.

> STATISTICAL MODEL FOR REPEATED MEASUREMENTS. A dataset consisting of values $x_1, x_2, \ldots, x_n$ of repeated measurements of the same quantity is modeled as the realization of a random sample $X_1, X_2, \ldots, X_n$. The model may include a partial specification of the probability distribution of each X_i.

- Pravděpodobnostní distribuce každého X_i se nazývá jako **modelová distribuce** (je to vlastně soubor všech distribucí pro každé X_i).
- **Modelovým parametrem** nazýváme konkrétní parametr **modelové distribuce**.
- Pokud není zaručena nezávislost nebo pravděpodobnostní distribuce každé proměnné náhodného výběru nejsou identické, pak musíme zvolit jiný statistický model.
- Máme-li statistický model pro náš statistický soubor, pak můžeme teprve usuzovat na vlastnosti modelové distribuce.

Statistický model

- Jaké vlastnosti **modelové distribuce** nás zajímají?
- Jaké vlastnosti modelové distribuce nejlépe odpovídají hledaným parametrům statistického souboru?
- Jaký **statistický soubor** musíme použít, abychom parametry modelové distribuce stanovili správně?
- Jaká modelová distribuce nejlépe pasuje na statistický soubor?
- Odpovědi nejsou triviální.
- Např. změřené doby trvání erupcí gejzírů (str. 25, přednáška 5) jsou náhodné výběry z nějaké neznámé *F*. Protože neznáme základní soubor měření, tak musíme stanovit kompletní křivku *F* ze změřeného náhodného výběru.
- Na druhou stranu pokud budeme mít náhodný výběr, který bude generován procesem se známou pravděpodobnostní distribucí (třeba exponenciální), tak *F* bude plně charakterizována, pokud stanovíme parametr λ této distribuce.

Statistický model

- Nebo se nemusíme zajímat o distribuci jako celek, ale jen o nějakou konkrétní vlastnost, číslo atp., které je třeba výsledkem našeho experimentu.
- Např.: provádíme měření *c* = číselná hodnota *c* + chyba měření.
- Tedy změřené *c* je neznámá konstanta a chyba měření je náhodná fluktuace.
- Pokud vyloučíme systematickou chybu měření, tak chyba měření může být modelována náhodnou proměnou s E[X] = 0 a Var[X] = konst.
- Pak měření *c* může být modelováno náhodným výběrem s neznámou E[X] a konečnou hodnotou Var[X]. Číslo *c* bude právě střední hodnota modelové distribuce určené na základě našeho statistického souboru.

Statistický model

- Nechť **statistický model** adekvátně popisuje naměřená data, potom **výběrový soubor** s odpovídajícími spočítanými charakteristikami (střed, MAD, medián atp.) by měl popisovat odpovídající číselné charakteristiky **modelové distribuce**.
- Př. pro velké *n* bude výběrový průměr (střed) blízko střední hodnotě modelové distribuce.
- Pokusme se nalézt srovnání mezi statistickým (výběrovým) souborem a vlastnostmi odpovídající pravděpodobnostní distribuce.

Empirická distribuční funkce

- Mějme náhodný výběr X_1, X_2, X_3,… z distribuce *F*. Potom $F_n(a)$ bude empirická distribuce náhodného výběru:

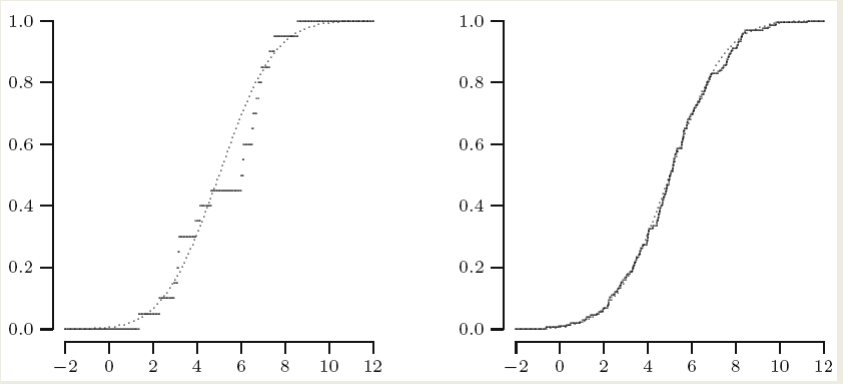
$$F_n(a) = \frac{\text{number of } X_i \text{ in } (-\infty, a]}{n}$$

- Ze zákona velkých čísel:

$$\lim_{n\to\infty} \mathrm{P}(|F_n(a) - F(a)| > \varepsilon) = 0.$$

$$F_n(a) \approx F(a).$$

- Př. empirická distribuční funkce normálního výběrového souboru pro *n* = 20 a 200.



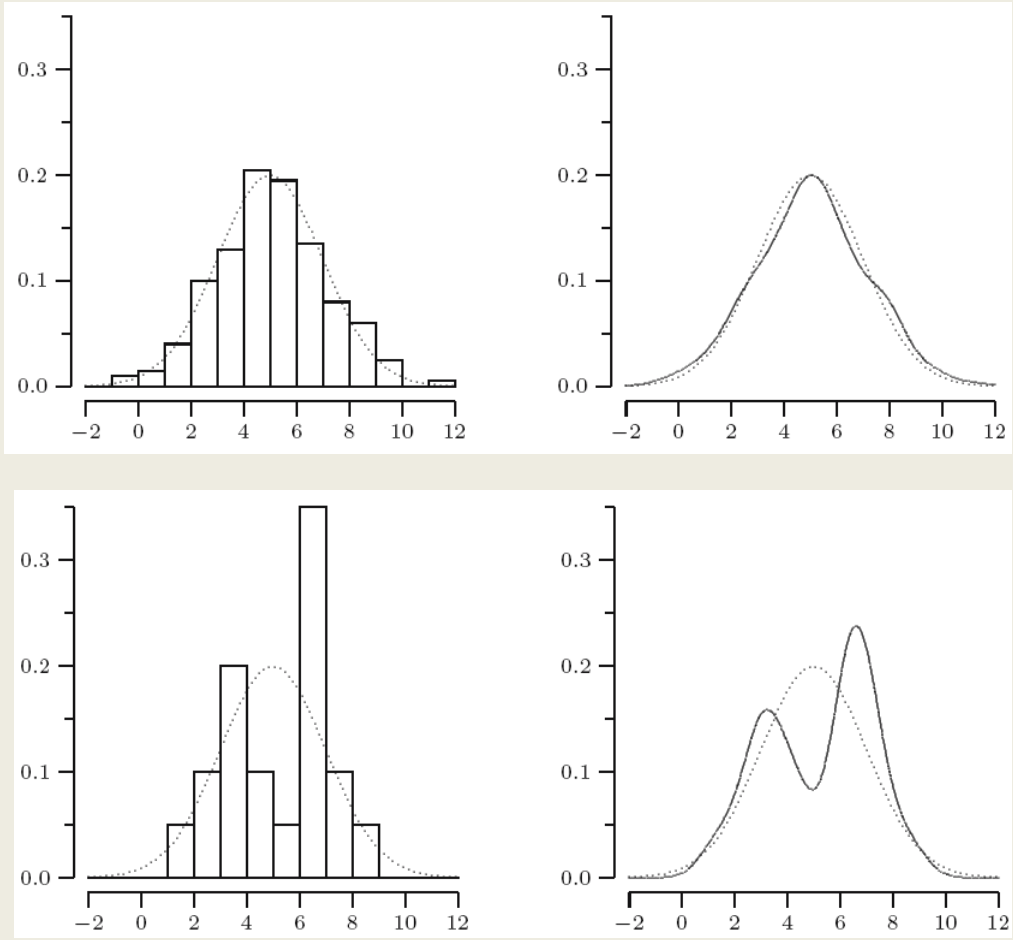
Histogram

- Máme náhodný výběr generovaný ze spojité distribuce s hustotou pravděpodobnosti *f*. Ze zákona velkých čísel plyne: $\frac{\text{number of } X_i \text{ in } (x-h, x+h]}{2hn} \approx f(x).$
- Pokud interval, v němž se X_i nachází je zároveň šířkou sloupce histogramu, potom výška histogramu je přibližně hodnotou *f* ve středu sloupce: $\text{height of the histogram on } (x-h, x+h] \approx f(x).$
- Podobně odhad jádrové hustoty náhodného výběru aproximuje hustotu pravděpodobnosti *f*

$$f_{n,h}(x) \approx f(x)$$

Histogram

- Histogram a odhad jádrové hustoty pro náhodný výběr generovaný z normálního rozdělení.
- Velikost počtu vzorků *n* značně ovlivňuje podobnost histogramu a původní hustoty pravděpodobnosti.
- *n* = 200 a 20.



Výběrový průměr, medián a empirické kvantily

- Pro normální distribuci *N*(5,4) platí, že E[X] = 5.
- Podle zákona velkých čísel výběrový průměr $\overline{X_n} \approx \mu$.
- Pro náhodný výběr generovaný z *N*(5,4) o velikosti 200 prvků dostaneme $\overline{X_n}$ = 5,012.
- Pro výběrový medián spočítáme Med(x_1, x_2, x_3,…, x_{200}) = 5.018.
- Vidíme, že výběrový medián mi dobře aproximuje medián *N*(5,4) distribuce.
- Obecně je možné ztotožnit empirický kvantil a *p* kvantil pravděpodobnostní distribuce: $q_n(p) \approx F^{\text{inv}}(p) = q_p$

Výběrový rozptyl a MAD

- Pro normální rozdělení $N(\mu, \sigma^2)$ je rozptyl Var [X] = σ^2 a směrodatná odchylka σ.
- Ze zákona velkých čísel platí: $S_n^2 \approx \sigma^2 \quad \text{and} \quad S_n \approx \sigma.$
- Pro konkrétní případ 200 prvkového statistického souboru generovaného z normálního rozdělení $N(5,4)$ dostaneme: $s_{200}^2 = 4.761 \quad \text{and} \quad s_{200} = 2.182$
- Pro medián absolutních odchylek (MAD) pro výše uvedený příklad dostaneme: MAD = 1,334. Značně se liší od σ!!!.
- Důvod: $\mathrm{MAD}(X_1, X_2, \ldots, X_n) \approx F^{\mathrm{inv}}(0.75) - F^{\mathrm{inv}}(0.5)$ pro distribuci symetrickou kolem mediánu. Pro $N(5,4)$ platí:

$$F^{\mathrm{inv}}(0.75) - F^{\mathrm{inv}}(0.5) = 2\Phi^{\mathrm{inv}}(0.75) = 1.3490$$

Statistický soubor vs. distribuce

- Shrnutí výběrových charakteristik a odpovídajících vlastností pravděpodobnostní distribuce, kterou aproximují.

Sample statistic	Distribution feature
Graphical	
Empirical distribution function F_n	Distribution function F
Kernel density estimate $f_{n,h}$ and histogram	Probability density f
(Number of X_i equal to a)/n	Probability mass function $p(a)$
Numerical	
Sample mean $\bar{X}_n$	Expectation μ
Sample median $\mathrm{Med}(X_1, X_2, \ldots, X_n)$	Median $q_{0.5} = F^{\mathrm{inv}}(0.5)$
pth empirical quantile $q_n(p)$	100pth percentile $q_p = F^{\mathrm{inv}}(p)$
Sample variance S_n^2	Variance σ^2
Sample standard deviation S_n	Standard deviation σ
$\mathrm{MAD}(X_1, X_2, \ldots, X_n)$	$F^{\mathrm{inv}}(0.75) - F^{\mathrm{inv}}(0.5)$, for symmetric F

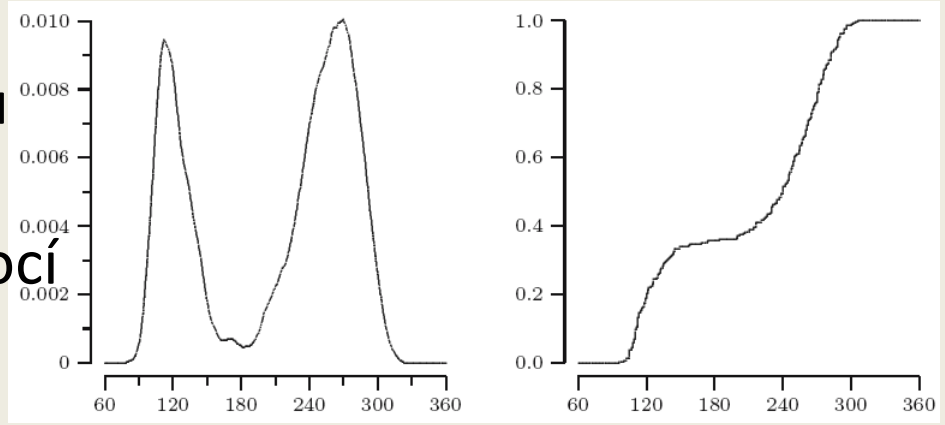
Vlastnosti „neznámých“ distribucí

- Viděli jsme, že pokud vygenerujeme náhodný statistický soubor z dané distribuční funkce, tak konkrétní vlastnosti této distribuce mohou být aproximovány z odpovídajících **výběrových charakteristik náhodného výběru**.
- V praxi se ale častěji setkáváme s opačnou situací: máme statistický soubor s *n* prvky, který je modelován jako realizace náhodného výběru s nějakou pravděpodobnostní distribucí, kterou ale neznáme.
- Tedy hledáme na základě našeho statistického souboru určité vlastnosti této neznáme pravděpodobnostní distribuce – typicky číselné charakteristiky.

Vlastnosti „neznámých“ distribucí

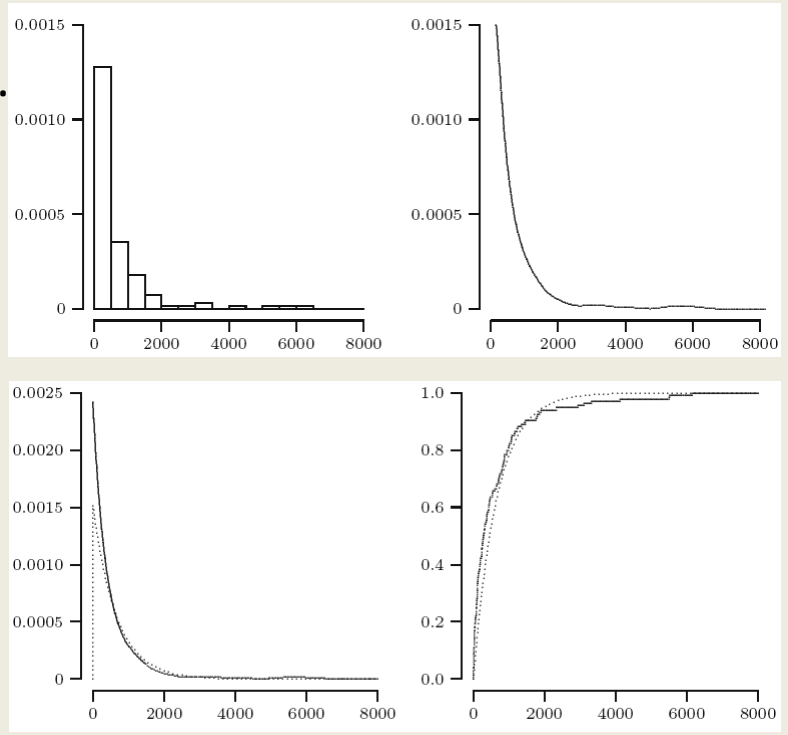
gejzír

- O fyzikálním pozadí toho moc nevíme. Nelze tedy specifikovat naměřený statistický soubor konkrétní parametrickou distribucí.
- Odhad jádrové hustoty a empirická distribuční funkce mi aproximují *f* a *F.*
- Je zřejmé, že to nejsou nějaké známé parametrické distribuce.
- Tedy odhadneme modelovou *f* pomocí odhadu jádrové hustoty a modelovou *F* pomocí empirické distribuční funkce.
- **Neparametrický odhad**.



Vlastnosti „neznámých“ distribucí *výpočetní chyba*

- Statistický soubor bude tvořit CPU čas mezi chybami vykonávaného počítačového kódu.
- Čas budeme modelovat jako realizace náhodného výběru z exponenciální distribuce.
- Uděláme si histogram a odhad jádrové hustoty – vypadá to jako exp(λ) rozdělení.
- Musíme stanovit parametr λ.
- E[X] = 1/λ a ze zákona velkých čísel bude $\lambda = 1/\bar{x} = 0{,}0015$.
- Porovnáme exp(0,0015) s grafickým zobrazením změřeného náhodného výběru – neparametrickým odhadem.
- Pokud je náš model správný, tak neparametrický odhad musí být více méně totožný s modelovou distribucí exp(0,0015).



- Data náhodného výběru jsou více akumulována kolem hodnoty 0 než předpokládá exp(0,0015).
- Je exp(λ) správný model statistického souboru?

Lineární regresní model

- Máme statistický soubor s dvěma náhodnými proměnnými – viz přednáška 5, str. 37.
- Z bodového grafu vidíme, že tvrdost dřeva je úměrná jeho hustotě: tvrdost = *g*(hustota).
- Je zde obsažena náhodnost, protože jsme naměřili pro jednu hodnotu hustoty více hodnot tvrdosti.
- Obecně takovou situaci modelujeme **regresním modelem**: tvrdost = *g*(hustota) + náhodná fluktuace.
- Jaký typ funkce *g* nejlépe pasuje na změřená data?
- Nic neznáme o fyzikálním pozadí; funkce může být skoro libovolná. S jistotou víme, že *g* bude rostoucí.
- Na první pohled můžeme říci, že lineárně rostoucí *g* bude dobře pasovat na naměřená data. Prvky náhodného výběru ležící mimo lineární fit budou modelovány náhodnou fluktuací vůči přímce.
- Tedy: tvrdost = $\alpha + \beta \cdot$(hustota dřeva) + náhodná fluktuace

Lineární regresní model

SIMPLE LINEAR REGRESSION MODEL. In a *simple linear regression model* for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we assume that $x_1, x_2, \ldots, x_n$ are nonrandom and that $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, n,$$

where $U_1, \ldots, U_n$ are *independent* random variables with $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = \sigma^2$.

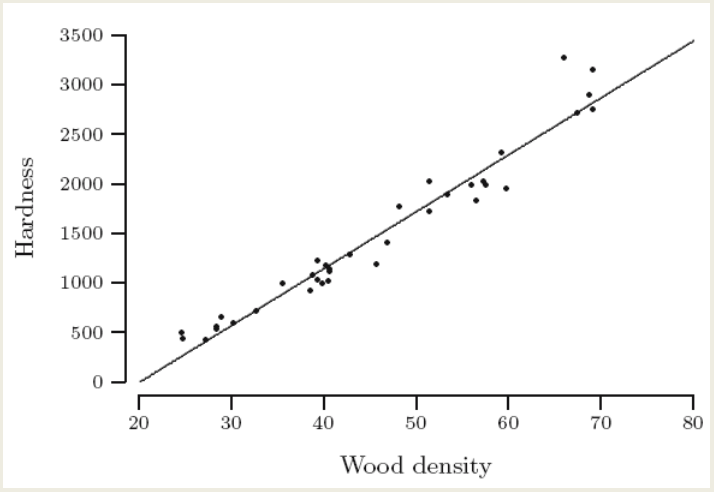
- Přímka $y = \alpha + \beta \cdot x$ se nazývá **regresní přímka**.
- Parametr α – úsek a β – směrnice regresní přímky.
- Proměnné *x* a *y* se nazývají také jako nezávislá a závislá proměnná.
- Náhodné proměnné U_1, U_2, U_3,… U_n jsou nezávislé, pokud se jednotlivá měření neovlivňují. Mají nulovou střední hodnotou a všechny stejný rozptyl.

Lineární regresní model

- Potom i Y_i jsou také nezávislé, ale Y_1, Y_2, Y_3,…, Y_n netvoří náhodný výběr.
- Protože každé Y_i má jinou střední hodnotu a tedy i jinou pravděpodobnostní distribuci.

$$\mathrm{E}[Y_i] = \mathrm{E}[\alpha + \beta x_i + U_i] = \alpha + \beta x_i + \mathrm{E}[U_i] = \alpha + \beta x_i$$

- Parametry α a β je třeba stanovit metodou nejmenších čtverců.
- Pak y = -1160,5 + 57,51·x



Bootstrap

Bootstrap – přesnost odhadu

- Víme jak stanovit parametry modelové distribuce ze znalostí výběrových charakteristik.
- Otázka:
  - jak moc se liší výběrové charakteristiky náhodného výběru od parametrů modelové distribuce?
  - jaká je pravděpodobnost, že výběrový průměr se bude lišit od E[X] o více jak ε?
- Potřebujeme znát, jak je výběrová charakteristika distribuována vzhledem k vlastnosti modelové distribuce. Např. zajímá nás pravděpodobnostní distribuce $\overline{X_n}$ - μ.
- Na příkladu doby erupce gejzíru vidíme, že pro daný statistický soubor pozorovaných časů $x_1, x_2, x_3, \ldots, x_n$ jde jen o jednu realizaci náhodného výběru $X_1, X_2, X_3, \ldots, X_n$.
- Pozorovaný výběrový průměr $\overline{x_n}$ je tedy určen jen z jedné realizace náhodné proměnné $\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$

Bootstrap pravidlo

- Mějme nový naměřený statistický soubor; odpovídající $\overline{x_n}$ je jiná realizace náhodné proměnné $\overline{X_n}$.
- Pokud chceme srovnávat výběrový průměr se střední hodnotou, tak musíme znát jak se mění realizace náhodné proměnné $\overline{X_n}$ – popíšeme to pravděpodobnostní distribucí $\overline{X_n}$.
- V principu distribuční funkci $\overline{X_n}$ lze stanovit z distribuční funkce *F* náhodného výběru. Ale obecně *F* neznáme!!!
- Musíme spočítat odhad $\hat{F}$ pro *F* a považovat náhodný výběr z $\hat{F}$ a odpovídající výběrový průměr jako náhradu náhodného výběru z *F* a odpovídajícího $\overline{X_n}$.
- Náhodný výběr z $\hat{F}$ = **bootstrap náhodný výběr** $X_1^*, X_2^*, \ldots, X_n^*$

Bootstrap pravidlo

- Boostrap výběrový průměr: $\bar{X}_n^* = \frac{X_1^* + X_2^* + \cdots + X_n^*}{n}$
- **Idea**: k aproximaci distribuce $\bar{X}_n$ použít distribuci $\bar{X}_n^*$.
- Tento postup nazýváme jako bootstrap pravidlo pro výběrový průměr:

> BOOTSTRAP PRINCIPLE. Use the dataset $x_1, x_2, \ldots, x_n$ to compute an estimate $\hat{F}$ for the "true" distribution function F. Replace the random sample $X_1, X_2, \ldots, X_n$ from F by a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, and approximate the probability distribution of $h(X_1, X_2, \ldots, X_n)$ by that of $h(X_1^*, X_2^*, \ldots, X_n^*)$.

- Pravidlo může být obecně aplikováno na jakoukoliv výběrovou charakteristiku aproximováním její pravděpodobnostní distribuce odpovídající bootstrap výběrovou charakteristikou.

Bootstrap pravidlo

- Jak dobře pravděpodobnostní distribuce $\bar{X}_n^*$ aproximuje distribuci $\bar{X}_n$?
- **Zobecnění**: jak dobře aproximuje distribuce zcela obecné bootstrap výběrové charakteristiky obecnou distribuci výběrové charakteristiky, o kterou se zajímáme?
- Bootstrap aproximaci můžeme vylepšit:
  - pokud využijeme centrovaného výběrového průměru ($\bar{X}_n$ - μ). Bootstrap verze bude: ($\bar{X}_n^*$ - μ*), kde μ* je střední hodnota distribuční funkce $\hat{F}$.
  - pokud obecně výběrovou charakteristiku „normalizujeme“ vzhledem k číselné charakteristice parametrické distribuce. Např. centrovaný výběrový medián Med(X_1, X_2, X_3,…, X_n) – $F^{inv}(0,5)$, nebo normalizovaný výběrový rozptyl S_n^2/σ^2.
- Nakonec musíme odpovědět na otázku jak aproximovat (odhadnout) $\hat{F}$ pro modelovou distribuci *F*?
- Pokud statistický soubor můžeme modelovat nějakou známou parametrickou distribucí např. *Exp*(λ) stačí nám aproximovat jen parametr *λ*, abychom mohli aproximovat $\hat{F}$.
- Tedy různé aproximace *F* nám tedy dávají různá pravidla pro bootstrap postupy.

Empirický bootstrap

- Mějme statistický soubor x_1, x_2, x_3,…, x_n jako realizaci náhodného výběru z diskrétní distribuce *F*, kdy každé x_i má pravděpodobnost 1/*n*.
- Nemůžeme nic říct o *F*. Nicméně můžeme stanovit empirickou distribuční funkci F_n daného statistického souboru:

$$\hat{F}(a) = F_n(a) = \frac{\text{number of } x_i \text{ less than or equal to } a}{n}.$$

- $\hat{F}$(a) nazýváme jako **empirický bootstrap**.
- Aplikujeme bootstrap pravidlo na centrovaný výběrový průměr. Tedy náhodný výběr X_1, X_2, X_3,…, X_n z *F* je nahrazen bootstrap náhodným výběrem $X_1^*, X_2^*, \ldots, X_n^*$ z F_n.
- Distribuce ($\bar{X}_n$ - μ) je aproximována ($\bar{X}_n^*$ - μ*), kde μ* je střední hodnota F_n.
- Jak dobrá je tato aproximace?
- Matematicky se dá dokázat, že bootstrap pravidlo dobře aproximuje centrovaný výběrový průměr!!!
- Ale např. pro normalizovanou výběrovou charakteristiku založenou na náhodném výběru z rov. rozdělení *U*(0,θ) empirický bootstrap nefunguje.

$$1 - \frac{\text{maximum of } X_1, X_2, \ldots, X_n}{\theta},$$

Empirický bootstrap

- Z předchozího plyne, že náhodná proměnná $\bar{X}_n^*$ má střední hodnotu:

$$\mu^* = \mathrm{E}[X_i^*] = x_1 \cdot \frac{1}{n} + x_2 \cdot \frac{1}{n} + \cdots + x_n \cdot \frac{1}{n} = \bar{x}_n.$$

- Tedy aplikace empirického bootstrap pravidla na ($\bar{X}_n$ - μ) znamená aproximace její distribuce distribucí ($\bar{X}_n^*$ - $\bar{x}_n$).
- Náhodná proměnná $\bar{X}_n^*$ je založená na náhodné proměnné X_i^*, jejíž distribuce je známá: nabývá hodnot x_1, x_2, x_3,…, x_n s pravděpodobností 1/*n*.
- Tedy můžeme stanovit ($\bar{X}_n^*$ - $\bar{x}_n$) a odpovídající pravděpodobnosti. Ale pro velké *n* je to trošku těžkopádná práce.

Empirický bootstrap

- Mnohem větší uplatnění najde metoda empirického bootstrap, když využijeme počítačové simulace.
- Analogicky k předchozímu příkladu budeme opakovaně generovat realizaci bootstrap náhodného výběru z F_n a vždy spočítáme odpovídající realizaci ($\bar{X}_n^* - \bar{x}_n$). Napočítané realizace nám dají dobrý přehled o distribuci náhodné proměnné ($\bar{X}_n^* - \bar{x}_n$).
- Realizace bootstrap náhodného výběru se nazývá jako **bootstrap statistický soubor**: $x_1^*, x_2^*, \ldots, x_n^*$
- Pro centrovaný výběrový průměr je simulační procedura následující:

Empirická bootstrap simulace

EMPIRICAL BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

corresponding to F_n.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

where

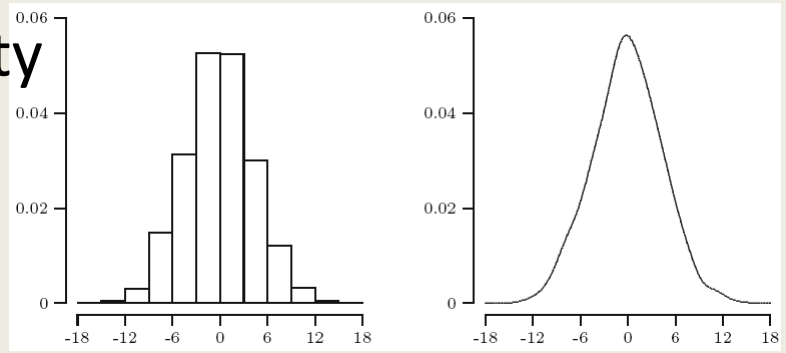
$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Generování hodnot x_i^* z F_n je ekvivalentní vybírání jednoho prvku z originálního statistického souboru x_1, x_2, x_3,…, x_n se stejnou pravděpodobností $1/n$.
- Obdobnou proceduru lze naformulovat na libovolnou jinou výběrovou charakteristiku.

Empirická bootstrap simulace

- Př.: aplikujme empirickou bootstrap simulaci na data doby pozorování erupcí gejzíru.
- Budeme simulovat centrovaný výběrový průměr pomocí 1000 simulací (opakování kroků 1 a 2).
- Histogram a odhad jádrové hustoty pro 1000 centrovaných bootstrap výběrových průměrů.



$$\bar{x}^*_{n,1} - \bar{x}_n \quad \bar{x}^*_{n,2} - \bar{x}_n \quad \cdots \quad \bar{x}^*_{n,1000} - \bar{x}_n.$$

- Dostali jsme tedy realizace náhodné proměnné ($\bar{X}_n^*$ - $\bar{x}_n$) a to mi odráží chování distribuce této proměnné. Tedy distribuce ($\bar{X}_n^*$ - $\bar{x}_n$) mi aproximuje distribuci ($\bar{X}_n$ - μ).

Empirická bootstrap simulace

- Předpokládejme, že střední hodnotu distribuce *F* modelující dobu trvání erupce gejzíru odhadneme číslem $\bar{x}_n$ = 209,3.
- Jak daleko je 209,3 od správné hodnoty μ?
- Protože statistický soubor je množina náhodně určených čísel, tak nemůžeme s absolutní přesností odpovědět na výše uvedenou otázku.
- Jediné co můžeme říct, je: s jakou pravděpodobností číslo 209,3 leží v dané vzdálenosti ($\mu - \varepsilon, \mu + \varepsilon$) od správné hodnoty μ.
- Tedy pokud máme statistický soubor o 272 prvcích a chceme získat představu jaká je hodnota μ, musíme si spočítat pravděpodobnost, že výběrový průměr se odchyluje od μ o více jak 5: $\mathrm{P}\left(|\bar{X}_n - \mu| > 5\right)$

Empirická bootstrap simulace

- Přímý výpočet pravděpodobnosti není možný, protože neznáme distribuci náhodné proměnné ($\bar{X}_n$ - μ).
- Protože distribuce ($\bar{X}_n^*$ - $\bar{x}_n$) aproximuje distribuci ($\bar{X}_n$ - μ), tak můžeme pravděpodobnost aproximovat následovně:

$$\mathrm{P}\left(|\bar{X}_n - \mu| > 5\right) \approx \mathrm{P}\left(|\bar{X}_n^* - \bar{x}_n| > 5\right) = \mathrm{P}\left(|\bar{X}_n^* - 209.3| > 5\right)$$

- Výpočet pravděpodobnosti $\mathrm{P}\left(|\bar{X}_n^* - 209.3| > 5\right)$ je zbytečně těžkopádný. Proto aproximujeme tuto pravděpodobnost tisící centrovanými bootstrap výběrovými průměry získanými pomocí bootstrap simulace:

$$\bar{x}_{n,1}^* - 209.3 \quad \bar{x}_{n,2}^* - 209.3 \quad \cdots \quad \bar{x}_{n,1000}^* - 209.3.$$

- Snadno nahlédneme, že přirozeným odhadem $\mathrm{P}\left(|\bar{X}_n^* - 209.3| > 5\right)$ je relativní četnost centrovaných bootstrap výběrových průměrů $\dfrac{\text{number of } i \text{ with } |\bar{x}_{n,i}^* - 209.3| \text{ greater than } 5}{1000}$

Empirická bootstrap simulace

- Na základě grafu na str. 30 lze pravděpodobnost vyčíslit:

$$\mathrm{P}\left(|\bar{X}_n - \mu| > 5\right) \approx \mathrm{P}\left(|\bar{X}_n^* - 209.3| > 5\right) \approx 0.227$$

- Aproximaci lze nekonečně zpřesňovat zvyšováním počtu opakování bootstrap postupu.

Parametrický bootstrap

- Mějme statistický soubor jakožto realizaci náhodného výběru s konkrétní distribuční funkce *F*.
- *F* je kompletně určena svým parametrem (nebo vektorem parametrů θ, pokud jich je více): $F = F_\theta$. Nemusíme tedy stanovit celou distribuci, ale stačí stanovit jen θ pomocí $\hat{\theta}$ a pak stanovit $\hat{F} = F_{\hat{\theta}}$.
- Takovéto bootstrap pravidlo nazýváme jako **parametrický bootstrap**.
- Opět chceme stanovit centrovaný výběrový průměr.
- Tedy střední hodnota F_θ musí být opět funkcí θ: $\mu = \mu_\theta$.
- Pravidlo parametrického bootstrap centrovaného výběrového průměru bude:

I. náhodný výběr $X_1, X_2, X_3, \ldots, X_n$ z F_θ je nahrazen bootstrap náhodným výběrem $X_1^*, X_2^*, \ldots, X_n^*$ z $F_{\hat{\theta}}$.

II. pravděpodobnostní distribuce $(\bar{X}_n - \mu_\theta)$ je aproximována $(\bar{X}_n^* - \mu^*)$. Střední hodnotu odpovídající $F_{\hat{\theta}}$ lze psát jako $\mu^* = \mu_{\hat{\theta}}$.

- Většinou parametrická bootstrap aproximace je mnohem lepší řešení než empirická bootstrap aproximace.

Parametrický bootstrap

- Výhoda spočívá v tom, že je možné v principu stanovit distribuci náhodné proměnné ($\bar{X}_n^* - \mu_{\hat{\theta}}$) přesně. Ale často to je těžkopádné řešení a počítání.
- Opět se vyplatí provést simulaci. Pro centrovaný výběrový průměr je následující:

PARAMETRIC BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, compute an estimate $\hat{\theta}$ for θ. Determine $F_{\hat{\theta}}$ as an estimate for F_θ, and compute the expectation $\mu^* = \mu_{\hat{\theta}}$ corresponding to $F_{\hat{\theta}}$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from $F_{\hat{\theta}}$.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \mu_{\hat{\theta}},$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

Parametrický bootstrap

- Jako příklad užijeme parametrickou bootstrap simulaci k zjištění, je-li exponenciální distribuce odpovídající model pro dobu mezi chybami v počítačovém kódu - viz str. 17.
- Víme, že exponenciální fit se zdál být rozumným modelem statistického souboru.
- Ke kvantifikování rozdílu mezi statistickým souborem a exponenciálním modelem vypočítáme maximální vzdálenost mezi empirickou distribuční funkcí F_n statistického souboru a exponenciální distribuční funkcí $F_{\hat{\lambda}}$ stanovenou ze statistického souboru:

$$t_{\mathrm{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\hat{\lambda}}(a)|$$

  kde $F_{\hat{\lambda}}$(a) = 0 pro a < 0, $F_{\hat{\lambda}}$(a) = 1-$e^{-\hat{\lambda}a}$ pro a $\geq$ 0 a $\hat{\lambda} = 1/\overline{x_n}$ je stanoveno ze změřeného statistického souboru.
- Statistický soubor hodnot t_{ks} nazýváme jako **Kolmogorov-Smirnovova** vzdálenost mezi F_n a $F_{\hat{\lambda}}$.

Parametrický bootstrap

- Jaká je podstata K-S vzdálenosti?
- Jestliže F popisuje reálnou matematickou distribuční funkci, potom empirická distribuční funkce F_n se musí více méně podobat F, ať už se F rovná nějaké exponenciální distribuci $F_\lambda = Exp(\lambda)$ nebo ne.
- Na druhou stranu pokud bude existovat reálná matematická distribuce F_λ potom odhad exponenciální distribuce $F_{\hat{\lambda}}$ bude podobný F_λ, protože a $\hat{\lambda} = 1/\overline{x_n}$ je blízko reálnému λ.
- Tedy
  - pokud $F = F_\lambda$ potom F_n a $F_{\hat{\lambda}}$ musí být blízko stejné distribuční funkci a tedy t_{ks} bude malé.
  - pokud $F \neq F_\lambda$ potom F_n a $F_{\hat{\lambda}}$ jsou každá blízko jiné distribuční funkci a tedy t_{ks} bude velké.
- t_{ks} vždy leží v intervalu mezi 0 a 1. Čím více se t_{ks} blíží k 1, je to indikace, že exponenciální model není správný.
- Pro příklad ze str. 17 platí: $\hat{\lambda} = 1/\overline{x_n} = 0{,}0015$ a $t_{ks} = 0{,}176$.

Parametrický bootstrap

- Tedy stanovená velikost K-S vzdálenosti t_{ks} = 0,176 pro zkoumaný příklad potvrzuje nebo vyvrací správnost exponenciálního modelu?
- Musíme zjistit: je-li pravda, že získaný datový soubor je opravdu realizací náhodného výběru z exponenciálního rozdělení F_λ, tak hodnota t_{ks} = 0,176 je neobvykle velká.
- Uvažujme výběrovou charakteristiku, která odpovídá t_{ks}.
- Odhad $\hat{\lambda} = 1/\overline{x_n}$ nahradíme náhodnou proměnnou $\hat{\Lambda} = 1/\overline{X_n}$ a empirickou distribuční funkci statistického souboru nahradíme empirickou distribuční funkcí náhodného výběru $X_1, X_2, X_3, \ldots, X_n$:

$$F_n(a) = \frac{\text{number of } X_i \text{ less than or equal to } a}{n}$$

- Potom t_{ks} je realizace výběrové charakteristiky: $T_{\mathrm{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\hat{\Lambda}}(a)|$

Parametrický bootstrap

- Abychom zjistili, že t_{ks} = 0,176 je mimořádně velká hodnota pro náhodnou proměnnou T_{ks} musíme stanovit distribuční funkci proměnné T_{ks}, ale to není možné protože neznáme λ.
- Budeme aproximovat distribuci T_{ks} pomocí parametrického bootstrap. Využijeme statistický soubor k určení λ pomocí $\hat{\lambda} = 1/\overline{x_n}$ = 0,0015.
- Náhodný výběr $X_1, X_2, X_3, \ldots, X_n$ z F_λ nahradíme bootstrap náhodným výběrem $X_1^*, X_2^*, X_3^*, \ldots, X_n^*$ z $F_{\hat{\lambda}}$. Pak budeme aproximovat distribuci T_{ks} její bootstrap verzí a spočítáme K-S vzdálenost pro T_{ks}:

$$T_{\mathrm{ks}}^* = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\Lambda}^*}(a)|$$

kde F_n^* je empirická distribuční funkce bootstrap náhodného výběru:

$$F_n^*(a) = \frac{\text{number of } X_i^* \text{ less than or equal to } a}{n}$$

Parametrický bootstrap

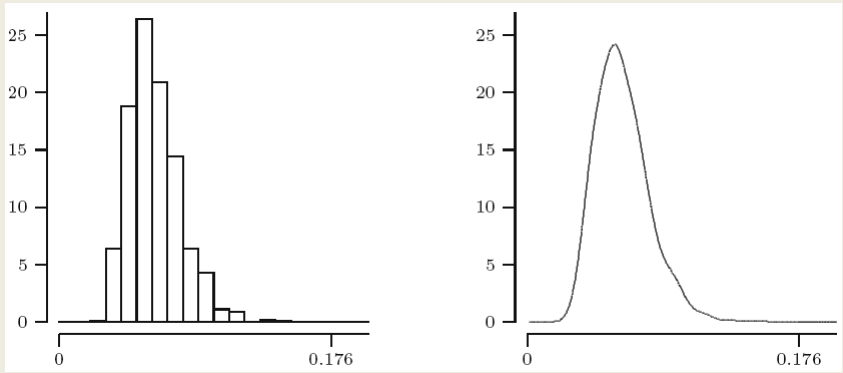
a $\widehat{\Lambda}^* = 1/\overline{X_n}^*$ s $\overline{X_n}^*$ jako průměrem bootstrap náhodného výběru.

- Je zřejmé, že bootstrap výběrová charakteristika T_{ks}^* je příliš složitá, abychom stanovili její pravděpodobnostní distribuci a proto provedeme parametrickou bootstrap simulaci:
  1. vygenerujeme bootstrap statistický soubor $x_1^*, x_2^*, x_3^*, \ldots, x_{135}^*$ z exponenciální distribuce s parametrem $\hat{\lambda} = 1/\overline{x_n} = 0{,}0015$.
  2. spočítáme bootstrap K-S vzdálenost: $t_{\mathrm{ks}}^* = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\lambda}^*}(a)|$

kde F_n^* je empirická distribuční funkce bootstrap statistického souboru a $F_{\widehat{\lambda}^*}$ je odhadovaná exponenciální distribuční funkce, kde $\hat{\lambda}^* = 1/\overline{x_n}^*$ je spočítané z bootstrap stat. souboru

Parametrický bootstrap

- Opakování simulace provedeme 1000 krát a dostaneme 1000 K-S vzdáleností, které zobrazíme jako histogram a odhad jádrové hustoty.
- Vidíme, že pokud dobu mezi chybami softwaru budeme modelovat exponenciální distribucí, tak K-S vzdálenost 0,176 je velmi nepravděpodobná.
- Tedy model pomocí exponenciální distribuce není správný.



- Důvod je ten, že Poissonův proces je špatný model pro sérii chyb, které při chodu programu nastávají.