Podstawy Przetwarzania Danych

Wykład 7: Próbkowanie danych

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

Wersja 1.0
5 marca 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

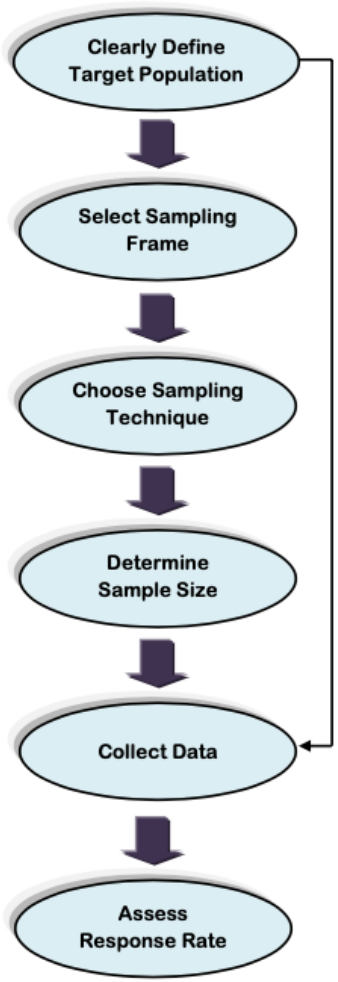
Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Próbkowanie

- Celem próbkowania jest zmniejszenie rozmiaru analizowanych danych.
- Redukcja jest realizowana poprzez tworzenie podzbioru analizowanych przypadków.
- Istotą próbkowania jest ustalenie rozmiaru pobieranej próbki i sposobu selekcji przypadków.

Proces próbkowania



1. Określenie rozmiaru populacji.
2. Wybór ram próbkowania.
3. Wybór technik próbkowania.
4. Określenie rozmiaru próbki.
5. Zebranie próbki.
6. Określenie współczynnika odpowiedzi.

Rysunek 1: Proces próbkowania [Taherdoost, 2016]

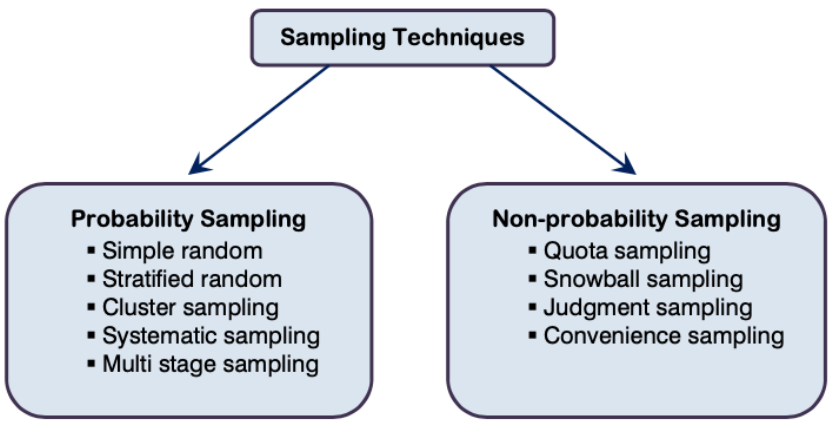
Określenie rozmiaru populacji

- Należy określić docelową populację, na temat której będziemy wyciągać wnioski.
- Nie należy utożsamiać populacji z dostępnym zestawem danych, zazwyczaj jest to większy, niedostępny zbiór.

Określenie ram próbkowania

- Ramy próbkowania to zakres danych z których będziemy pobierać próbki.
- Zazwyczaj jest to dostępny zbiór danych, na który możemy nałożyć dodatkowe ograniczenia.
- Ramy próbkowania powinny być reprezentatywne dla całej populacji.

Wybór techniki próbkowania



Rysunek 2: Techniki próbkowania [Taherdoost, 2016]

- Techniki próbkowania dzielą się na
  - Probabilistyczne.
  - Nieprobabilistyczne.
- Tylko techniki probabilistyczne pozwalają na wyciągnięcie statystycznie istotnych wniosków dla całej populacji.
- Jednakże próbkowanie nieprobabilistyczne pozwala zaprojektować badania, które pogłębią nasze zrozumienie przedmiotu analizy.

Określenie rozmiaru próbki

- W celu zapewnienia możliwości generalizacji wniosków uzyskanych na losowej próbce danych i uniknięciu tendencyjnych wyników należy zapewnić jej odpowiedni rozmiar.
- Rozmiar próbki zależy od kilku czynników, które mogą wydawać się nieintuicyjne, gdyż nie odnoszą się one tylko do proporcjonalności próbki.
- Czynniki, które należy uwzględnić
  - Bezwzględny rozmiar próbki względem złożoności populacji.
  - Cele badania.
  - Zastosowane metody statystyczne.

Zbieranie danych a współczynnik odpowiedzi

- Określając rozmiar próbki należy brać pod uwagę współczynnik odpowiedzi.
- Współczynnik odpowiedzi określa jaki procent danych z próbki nadaje się do wykorzystania.
- Jego wartość jest zależna od braków w danych.
- W praktyce ustala się, że zbierana próbka będzie większa niż estymowany rozmiar, aby pokryć elementy wyeliminowane przez współczynnik odpowiedzi.

Próbkowanie probabilistyczne

- W próbkowaniu probabilistycznym każdy obiekt z populacji na równe szanse na włączenie do próbki.
- Ogranicza tendencyjność.
- Jednakże osiągnięcie zakładanego poziomu błędu może być bardzo kosztowne pod kątem budowania próbki.

Próbkowanie losowe

- Każda obserwacja z populacji ma równą szansę na trafienie do próbki.
- Podejście jest proste i nie wymaga dodatkowych analiz.
- Problemy związane z tym podejściem:
  - Ramy próbkowania muszą uwzględniać całą populację.
  - Koszt zdobycia danych może być bardzo wysoki.
  - nie ma gwarancji reprezentatywności.

Próbkowanie systematyczne

- W próbkowaniu systematycznym rozpatrujemy każdy n-ty przypadek poczynając od losowego początku.
- Metoda jest prosta i czasami łatwiejsza w implementacji niż próbkowanie losowe.
- Może obniżać reprezentatywność.

Próbkowanie stratyfikacyjne

- W próbkowaniu stratyfikacyjnym populacja jest dzielona na straty (podgrupy).
- Próbka losowa jest pobierana z każdej podgrupy.
- Podgrupa odpowiada naturalnym cechom obiektów.
- Podgrupa może odpowiadać
  - rozmiarowi przedsiębiorstwa,
  - płci,
  - zawodowi itp..
- Próbkowanie stratyfikacyjne jest stosowane gdy populacja charakteryzuje się dużą zmiennością.
- Podejście zapewnia reprezentację wszystkich podgrup, ale dobór podgrup może być problematyczny.

Próbkowanie grupowe

- Próbkowanie grupowe jest stosowane wtedy, gdy można uprzednio podzielić dane na rozłączne grup.
- Grupy mogą być reprezentowane przez
  - obszary geograficzne,
  - branże itp..
- Dla każdej grupy z osobna wybiera się próbkę losową.
- Wyniki uzyskane z różnych grup mogą być trudne do interpretacji.

Próbkowanie wieloetapowe

- Próbkowanie wieloetapowe polega na wstępnym podziale populacji na podgrupy.
- Następnie losowo wybieramy podgrupy, które będą analizowane.
- W kolejnych krokach można podzielić grupy na podgrupy i wybrać ich losowych przedstawicieli.
- Z finalnie wybranych grup losuje się próbkę do analizy.
- Przykładowo można wylosować województwa, spośród których wylosujemy gminy, spośród których wylosujemy miasta z których zbierzemy próbki do badań.
- Podejście ogranicza zakres koniecznych badań.

Próbkowanie nieprobabilistyczne

- Próbkowanie nieprobabilistyczne jest często powiązane z badaniem przypadku i badaniami jakościowymi.
- Badania skupiają się na małej grupie obserwacji, aby analizować konkretny przypadek nie mający statystycznego przełożenia na całość populacji.
- Badana próbka nie musi być reprezentatywna ani losowa ale dobór obserwacji musi być racjonalny.

Próbkowanie kwotowe

- Próbkowanie kwotowe dobiera obserwacje tak, aby oddawały one charakter całości populacji.
- W próbce chcemy osiągnąć taką samą dystrybucję charakterystyk jaką ma ogół populacji.
- Dobór nie odbywa się w sposób losowy.
- Uzyskujemy kontrolowaną próbkę, ale możliwa jest tendencyjność wynikająca z kryterium doboru obserwacji.

Próbkowanie lawinowe

- Próbkowanie lawinowe jest metodą, która używa dotychczasowych obiektów obserwacji do zwiększania próbki badawczej.
- Stosowane w małych populacjach, gdzie uzyskanie nowych obiektów obserwacji może być trudne ze względu na ich zamkniętą naturę.
- Mocno tendencyjne i czasochłonne, ale może być jedyną możliwością pozyskania próbki w pewnych okolicznościach.

Próbkowanie dogodne

- W próbkowaniu dogodnym wybieramy obiekty do próbki ze względu na łatwość ich doboru.
- Najczęściej stosowane w badaniach prowadzonych przez studentów, ze względu na niski koszt pozyskania danych.
- Może dotyczyć i ograniczać się do badań na rodzinie i znajomych lub zjawisk obserwowanych przez badacza w jego codziennym życiu.
- Pozyskana próbka nie jest reprezentatywna i może być tendencyjna.
- Nierekomendowane w normalnych badaniach naukowych.

Próbkowanie krytyczne

- Próbkowanie krytyczne ma miejsce gdy badacz włącza do próbki obserwacje według własnego wyboru.
- Zazwyczaj bazuje ono na przekonaniu, że wybrane obserwacje są krytyczne dla badanego zjawiska.
- Mocno tendencyjne i z dużym prawdopodobieństwem niereprezentatywne podejście.
- Przydatne przy eksploracji danych, gdy chcemy zweryfikować własny wstępny osąd o badanym zjawisku.

Znaczenie rozmiaru próbki

- Większa próbka zmniejsza prawdopodobieństwo otrzymania tendencyjnych wyników.
- Jednakże koszt poprawy estymacji nie rośnie liniowo i poprawianie jakości staje się coraz bardziej kosztowne.
- Rozmiar próbki jest wyznaczany w zależności od zadania, które przed nami stoi.

Ustalanie udziału w populacji

- Jeżeli chcemy oszacować udział obserwacji z pewną własnością w populacji to możemy oszacować jej częstość występowania jako

$$\hat{p} = \frac{X}{n},$$

gdzie X liczba elementów z daną własnością w populacji o rozmiarze n.

- Jeżeli obserwacje są niezależne to estymator będzie miał rozkład dwumianowy.
- Ponieważ nie znamy rozkładu poszukiwanych obserwacji w próbce możemy założyć, że $p = 0.5$ co zmaksymalizuje wariancję.

Określenie przedziału ufności

- Dla odpowiednio dużego n rozkład będzie przybliżał rozkład normalny.
- Wtedy przedział ufności wyniesie:

$$\left(\hat{p} - Z\sqrt{\frac{0.25}{n}}, \hat{p} + Z\sqrt{\frac{0.25}{n}}\right)$$

gdzie Z to wartość Z-testu dla zakładanego poziomu ufności, a 0.25 jest wynikiem maksymalizacji wyrażenia $p(1-p)$.

- Dopuszczając margines błędu E otrzymujemy

$$E = Z\sqrt{\frac{0.25}{n}} \implies n = 0.25\frac{Z^2}{E^2}$$

- Jest to oszacowanie dla najgorszego przypadku, w ogólności

$$n = p(1-p)\frac{Z^2}{E^2}.$$

Wyliczanie rozmiaru próbki dla danych nominalnych

- Rozmiar próbki dla danych nominalnych możemy wyliczyć jako

$$n = p(1-p)\frac{Z^2}{E^2}$$

gdzie

- n wymagany rozmiar próbki,
- p procent wystąpień wartości nominalnej,
- E dopuszczalny margines błędu,
- z wartości odpowiadająca poziomowi ufności.

Margines błędu

- Współczynnik E określa margines błędu (poziom precyzji) lub ryzyko, które badacz godzi się przyjąć.
- Zakładając błąd $E = 0.5$ godzimy się z faktem, że uzyskany wynik x odnosi się tak naprawdę do przedziału $x \pm 0.5$
- Współczynnik wpływa znacząco na rozmiar próbki, którą należy przebadać.
- Można przyjąć następującą zależność

$$E \sim \frac{1}{\sqrt{n}}$$

Poziom ufności

- Poziom ufności określa jaki procent uzyskanych obserwacji mieści się w przedziale zdefiniowanym przez E
- Poziom ufności rzędu 95 procent oznacza, że 95 na 100 obserwacji x będzie rzeczywiście mieściło się w przedziale $x \pm E$
- Poziom ufności jest określany statystyką Z.

Procent wystąpień

- Parametr p określa procent przedstawicieli danej kategorii w populacji.
- W przypadku cech opisywanych wieloma kategoriami należy zastosować podejście jeden kontra wiele.
- W przypadku określania rozmiaru próbki przed rozpoczęciem procesu zbierania danych należy założyć, najkosztowniejszy wariant $p = 0.5$.

Dobór próbki w praktyce

`	Variance of the population P=50%					
	Confidence level=95% Margin of error			*Confidence level=99% Margin of error*		
Population Size	5	3	1	5	3	1
50	44	48	50	46	49	50
75	63	70	74	67	72	75
100	79	91	99	87	95	99
150	108	132	148	122	139	149
200	132	168	196	154	180	198
250	151	203	244	181	220	246
300	168	234	291	206	258	295
400	196	291	384	249	328	391
500	217	340	475	285	393	485
600	234	384	565	314	452	579
700	248	423	652	340	507	672
800	260	457	738	362	557	763
1000	278	516	906	398	647	943
1500	306	624	1297	459	825	1375
2000	322	696	1655	497	957	1784
3000	341	787	2286	541	1138	2539
5000	357	879	3288	583	1342	3838
10000	370	964	4899	620	1550	6228
25000	378	1023	6939	643	1709	9944
50000	381	1045	8057	652	1770	12413
100000	383	1056	8762	656	1802	14172
250000	384	1063	9249	659	1821	15489
500000	384	1065	9423	660	1828	15984
1000000	384	1066	9513	660	1831	16244

Rysunek 3: Parametryczny rozmiar próbki [Taherdoost, 2016]

Ustalanie wartości średniej dla populacji

- Chcemy oszacować wartość średnią zmiennej o wariancji σ dla całej populacji n elementów.
- Błąd standardowy dla zmiennej wynosi $\frac{\sigma}{\sqrt{n}}$ więc dla rozkładu normalnego możemy określić przedział ufności jako

$$\left(\bar{x} - \frac{Z\sigma}{\sqrt{n}}, \quad \bar{x} + \frac{Z\sigma}{\sqrt{n}}\right).$$

- Dopuszczając margines błędu E otrzymujemy

$$E = \frac{Z\sigma}{\sqrt{n}} \implies n = \frac{\sigma^2 Z^2}{E^2}$$

Wielkości próby dla zmiennej nienormatywnej

- Wyliczenie wielkości próby dla zmiennej nienormatywnej można przeprowadzić według wzoru

$$n = \frac{\sigma^2 Z^2}{E^2}$$

  - σ - odchylenie standardowe
  - Z - poziom ufności
  - E - margines błędu
- Wynik zawsze zaokrąglamy w górę.

Bibliografia I

[Taherdoost, 2016] Taherdoost, H. (2016).

Sampling methods in research methodology; how to choose a sampling technique for research.

International Journal of Academic Research in Management, 5(2):18–27.