# Statystyka i analiza danych - laboratorium 12 - Podsumowanie, prawidłowy wybór testu

Paweł Misiorek

Instytut Informatyki
Politechnika Poznańska (PP)
Piotrowo 3, 60-965 Poznan, Poland
Email: pawel.misiorek@put.poznan.pl

13/16 czerwca 2023

# Zakres podsumowania

Zagadnienia 2 części semestru:

- Test t
- Testy dwóch zbiorowości
- Korelacja i regresja
- Analiza regresji
- Test $\chi^2$
- Testy nieparametryczne

# Test t

- Kiedy trzeba używać: próba jest mała ($n < 30$), a wariancja nieznana.
- Wariancje estymujemy:

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

- Statystyka t ma postać:

$$t_n = \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{X}_n - \mu}{s}\sqrt{n} \approx t(n-1)$$

- wykonując test korzystamy z tablic rozkładu $t(n-1)$, tj. rozkładu t-studenta z $n-1$ stopniami swobody
- dla $n \geq 30$ rozkład normalny i rozkład t-studenta są nieomal tożsame

# Testy dwóch zbiorowości

- Test sparowany (statystyka - ustandaryzowana średnia różnic); test T (rozkład $t(n-1)$, dla $n \geq 30$ można przybliżać rozkładem $N(0,1)$, gdzie $n$ to liczba par)
- Test niesparowany, duża próba (statystyka - ustandaryzowana różnica średnich; test Z, wariancje znane albo estymowane z danych)
- Test niesparowany, mała próba - test T - wersja z równymi wariancjami, estymator wariancji łącznej
- Test na równość wariancji - test F z użyciem rozkładu Fishera-Snedecora

# Korelacja i regresja

- Kowariancja, korelacja Pearsona - wzory ogólne (dla populacji):
  - Kowariancja:
$$\sigma_{X,Y} = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right)$$
  - Korelacja (współczynnik korelacji Pearsona):
$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$
- Kowariancja, korelacja Pearsona - estymatory (dla próby):
  - Kowariancja:
$$S_{X,Y} = \frac{1}{n-1}\left((X - \overline{X})(Y - \overline{Y})\right)$$
  - Współczynnik korelacji Pearsona:
$$r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y}$$

# Korelacja i regresja - przypomnienie

- Test na istotność korelacji - test T (o $n-2$ stopniach swobody):
  - dla $H_0$: $\rho = 0$ i $H_1$: $\rho \neq 0$ (lub $\rho > 0$ lub $\rho < 0$)
  - dla statystyki zdefiniowanej jako:

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

- Regresja liniowa jednej zmiennej $Y = \beta_0 + \beta_1 X$:
  - Wyznaczanie współczynników (metoda najmniejszych kwadratów):

$$\hat{\beta}_1 = \frac{S_{X,Y}}{S_X^2} = r_{X,Y}\frac{S_Y}{S_X}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

# Analiza regresji

- Model regresji - regresja liniowa z szumem homoskedastycznym (ze stałą i niezależną wariancją)
- SST, SSR, SSE
- Współczynnik determinacji

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Test na istotność regresji (rozkład F Snedecora) - globalny test na istotność regresji wielorakiej
- Test na istotność pojedynczego parametru (test T)

# Analiza regresji

- Test na istotność regresji (test F):
  - W globalbym teście na istotność regresji liniowej (w tym regresji wielorakiej) korzysta się ze statystyki

$$F = \frac{SSR/k}{SSE/(n-(k+1))}$$

    oraz z rozkładu F Snedecora z $k$ i $n-k-1$ stopniami swobody, gdzie $k$ to liczba zmiennych wyjaśniających, a $n$ to liczba obserwacji dla poszczególnych zmiennych.
  - Uwaga 1: dla $k = 1$ test ten równoważny testowi T na istotność korelacji
  - Uwaga 2: współczynnik $\beta_0$ nie wchodzi do układu hipotez.

# Analiza regresji

- Test na istotność pojedynczego parametru $\beta_i$ (test T):
  - W teście na istotność pojedynczego współczynnika liniowego dla regresji wielorakiej korzysta się ze statystyki

$$T = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

    i rozkładu t-studenta o $n - k - 1$ stopni swobody ($k$ - liczba zmiennych wyjaśniających).
  - W przypadku regresji liniowej jednej zmiennej sprowadza się to do rozkładu t-studenta z $n - 2$ stopniami swobody - test ten de facto wtedy jest tożsamy na istotności korelacji Pearsona między badaną zmienną wyjaśniającą $X_i$, a zmienną wyjaśnianą $Y$.

# Rozkład $\chi^2$

- Definicja i podstawowe cechy rozkładu $\chi^2$ (suma kwadratów zmiennych o rozkładzie normalnym ustandaryzowanym - coś zawsze nieujemnego)
- Zastosowania do analizy danych jakościowych (nienumerycznych)
  - Test $\chi^2$ na zgodność rozkładu ($k-1$ stopni swobody, $k$ - liczba możliwych wartości zmiennej skategoryzowanej)
  - Test $\chi^2$ na niezależność zmiennnych ($(k-1)\cdot(w-1)$ stopni swobody, $k$,$w$ - liczba kolumn i wierszy)
  - wspólny klucz - zdefiniowanie statystyki bazującej na sumie kwadratów odchyleń
  - umiejętność obliczania stopni swobody (liczba źródeł błędu - źródeł narastania statystyki)

# Miary zależności zmiennych nominalnych

- współczynnik Yule'a - $\phi$

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

gdzie: $\chi^2$ - wartość statystyki $\chi^2$ dla testu na niezależność zmiennych dla $n$ obserwacji

- współczynnik kontyngencji C Pearsona

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- współczynnik V Crammera

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, w-1)}}$$

gdzie $k$ - liczba kolumn, $w$ - liczba wierszy

# Testy nieparametryczne (1/2)

- Test U Manna-Whitneya (zwany też testem Wilcoxona dla sum rang)
- Operacja rangowania
- Test znaków (odpowiednik testu sparowanego T dla dwóch populacji)
- Sparowany test Wilcoxona (odpowiednik testu sparowanego T dla dwóch populacji)
- Korelacja Spearmana

# Testy nieparametryczne (2/2)

- Testy nieparametryczne (brak założeń co do rozkładu normalnego)
- Nieparametryczne odpowiedniki testów sparowanych
  - Test znaków, statystyka $Z = \frac{2T-n}{\sqrt{n}}$ oraz tablice rozkładu normalnego
  - Sparowany test Wilcoxona (Wilcoxon signed-rank test), statystyka: odpowiednia suma rang (uwaga ma on własne tablice)
- Nieparametryczny odpowiednik testu niesparowanego
  - Test Manna-Whitneya (zwany też testem Wilcoxona dla sum rang) - albo testem MWW

# 2 część semestru z perspektywy testów

- Test t
- Testy parametryczne (dwóch zbiorowości)
- Test F
- Test $\chi^2$
- Testy na istotność
- Testy nieparametryczne (znaków, Manna-Whitneya (MWW), Wilcoxona)

# Testy parametryczne dwóch zbiorowości

- Testy wartości średniej (zmienne liczbowe)
- Próby sparowane (czasem zwane “zależnymi”) vs próby niesparowane (“niezależne”)
- Kluczowa liczność prób - wybór miedzy testem T a Z ($n \geq 30$), ale pamietajmy, że:
  - gdy wiemy, że rozkład jest normalny i znane jest odchylenie standardowe dla całej populacji wykonuje się test Z także dla małych prób
  - z kolei gdy nieznane jest odchylenie standardowe zawsze wykonuje się test T z użyciem estymatora $s$ dla $\sigma$ (a w przypadku dużej populacji test T aproksymuje się testem Z)
  - dla małej populacji konieczne jest spełnienie założenie o rozkładzie normalnym (zarówno dla testu T jak i (tym bardziej) dla testu Z)

# Testy parametryczne dwóch zbiorowości

- Próby niesparowane - wybór między testem T a Z
  - Test Z wybieramy gdy próba duża ($n \geq 30$) - rozkłady można przybliżać rozkładem normalnym, a w przypadku gdy nieznane są wariancje estymujemy je (wtedy test Z jest wybrany jako aproksymacja testu T)
  - Z kolei gdy próba jest mała z powodu "braku zaufania do estymatorów wariancji" nie można użyć testu Z - wykonuje się test T - estymatory z prób zwykle stosuje się do policzenia tzw. wariancji łącznej (zakłada się równość wariancji w obu populacjach - oczywiście ta wariancja jest nieznana) - liczba stopni swobody to $n_1 + n_2 - 2$.

# Gdzie jeszcze stosujemy test T

- Do zbadania istotności korelacji Pearsona ($n-2$ stopnie swobody, $n$ - liczba obserwacji)
- Do zbadania istotności korelacji Spearmana ($n-2$ stopnie swobody)
- Do zbadania istotności danego współczynnika regresji wielorakiej ($n-m-1$ stopnie swobody, $m$ - liczba zmiennych objaśniających - w przypadku regresji liniowej $n-2$ (de facto wtedy jest to test na istotności korelacji Pearsona))
- Uwagi:
  - ważne by dobrze zdefiniować statystykę i sformułować hipotezy oraz pamiętać o założeniach
  - dla dużych $n$ można wartości z tablic rozkladu $T$ zastąpić tymi z rozkładu normalnego

# Test F

- Wiele wynika z samej definicji statystyki F jako ilorazu dwóch statystyk o rozkładzie $\chi^2$ (w ogólności dodatnich wartości, czyli odchyleń, wariancji, błędów)
- Ma on dwa parametry odpowiadające stopniom swobody dla obu dzielonych przez siebie zmiennych
- Test F stosuje się zatem:
  - W teście na zgodność wariancji w dwóch populacjach (statystyką jest oczywiście iloraz estymatorów wariancji, a stopni swobody mamy $n_1 - 1$ oraz $n_2 - 1$; $n_1$, $n_2$ - liczności w próbach,
  - W globalbym teście na istotność regresji wielorakiej gdzie statystyką jest

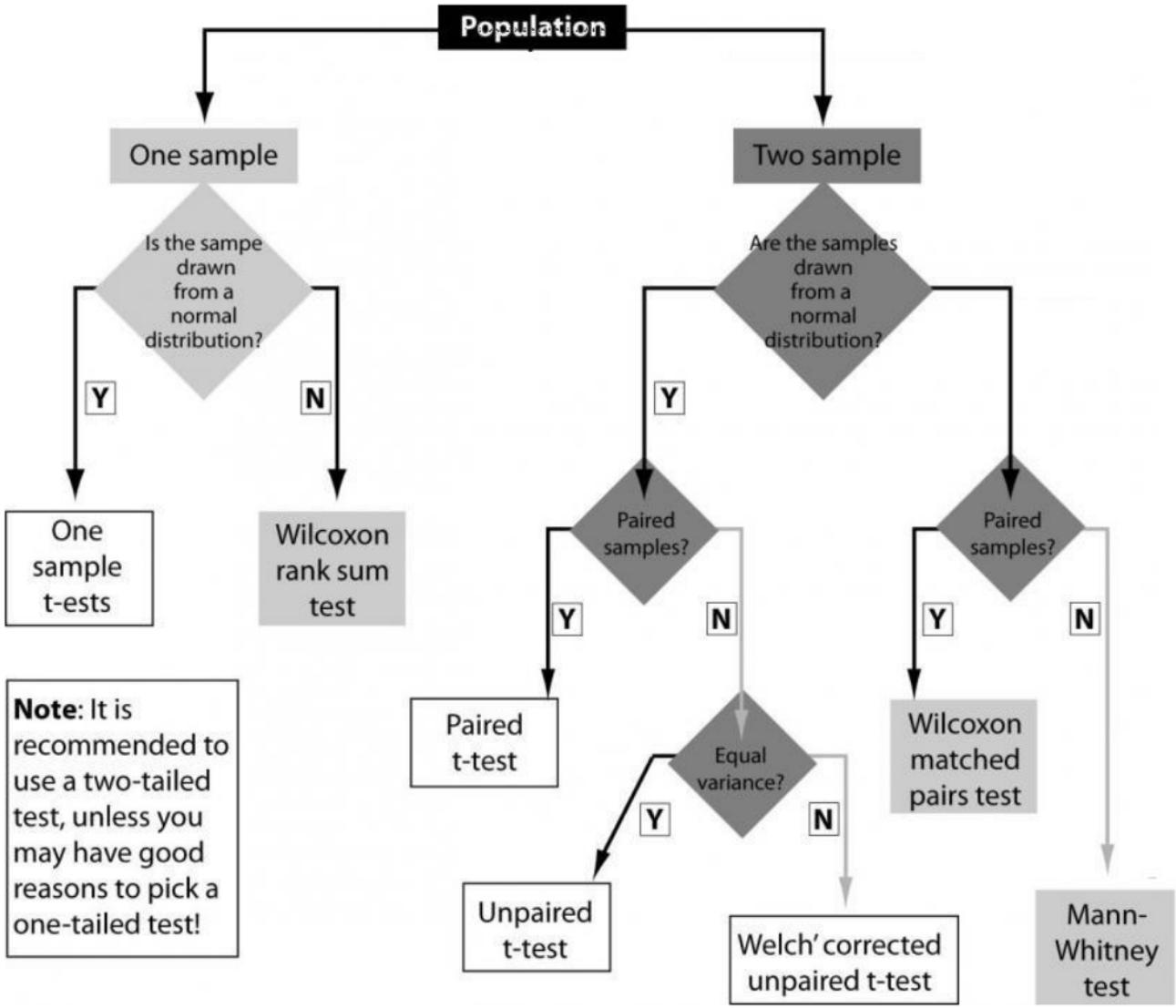
$$\frac{SSR/k}{SSE/(n-(k-1))}$$

i korzysta sie z rozkładu F Snedecora z $k$ i $n-k-1$ stopniami swobody (dla $k = 1$test ten równoważny testowi T na istotność korelacji) - uwaga: współczynnik $\beta_0$ nie wchodzi do układu hipotez.

# Wybór testu

- Testy nieparametryczne uznaje się za “słabsze” (o mniejszej mocy) od ich parametrycznych odpowiedników ze względu na brak założeń co do rozkładu normalnego - w zakresie rozkładów dla estymatorów w próbie
- Wg wykładu metody nieparametryczne:
  - albo nie zajmują się parametrami populacji takimi jak $\mu$ albo $\sigma$,
  - albo nie wymagają konkretnych rozkładów populacji (klasycznie: nie wymagają rozkładu normalnego)
- Stosujemy je także gdy zachodzi niezgodność wariancji albo dane zdefiniowane są na skali porządkowej
- Jak zwiększyć moc testu?
  - zwiększ liczebność próby
  - zmniejsz losowość (wariancje) w próbie
  - wg Wikipedii: *wykorzystaj testy lub struktury badania, które mają z natury wyższą moc statystyczną, takie jak metody bayesowskie,* **testy parametryczne**, *prerejestrowane testy hipotez jednostronnych, oraz badania w schemacie wewnątrzgrupowym (generalnie lepiej zwiększać liczbę uczestników niż powtórzeń)*

# Wybór testu




Rysunek: [źródło] Materiały M.Lango

# Pytania

Pytania proszę zadawać:

- w czasie konsultacji (zoom),
- na Slacku,
- albo mailowo.