Cele analiz przejściowych (interim analyses)

Wcześniejsze przerwanie (early stopping) próby z powodu

- bezpieczeństwa (safety)
- skuteczności (efficacy)
- nieskuteczności (futility)

Modyfikacja układu doświadczalnego na podstawie zaobserwowanych wyników mająca na celu

- play the winner / drop the loser
- utrzymanie mocy statystycznej
- *dowolną* modyfikację, niezależnie od przyczyny lub danych, z kontrolą p-stwa błędu I rodzaju α

Metodologia analiz przejściowych

Wielostopniowe układy dośw. (multi-stage designs)

Układy „bezszwowe” (seamless designs) np. fazy II/III

Układy sekwencyjne (sequential designs)

Układy z sekwencyjnymi grupami (group-sequential)

Korekta liczebności próbki (sample size adjustments)

Stochastyczne „obcinanie” (stochastic curtailment)

Układy adaptacyjne (adaptive designs)

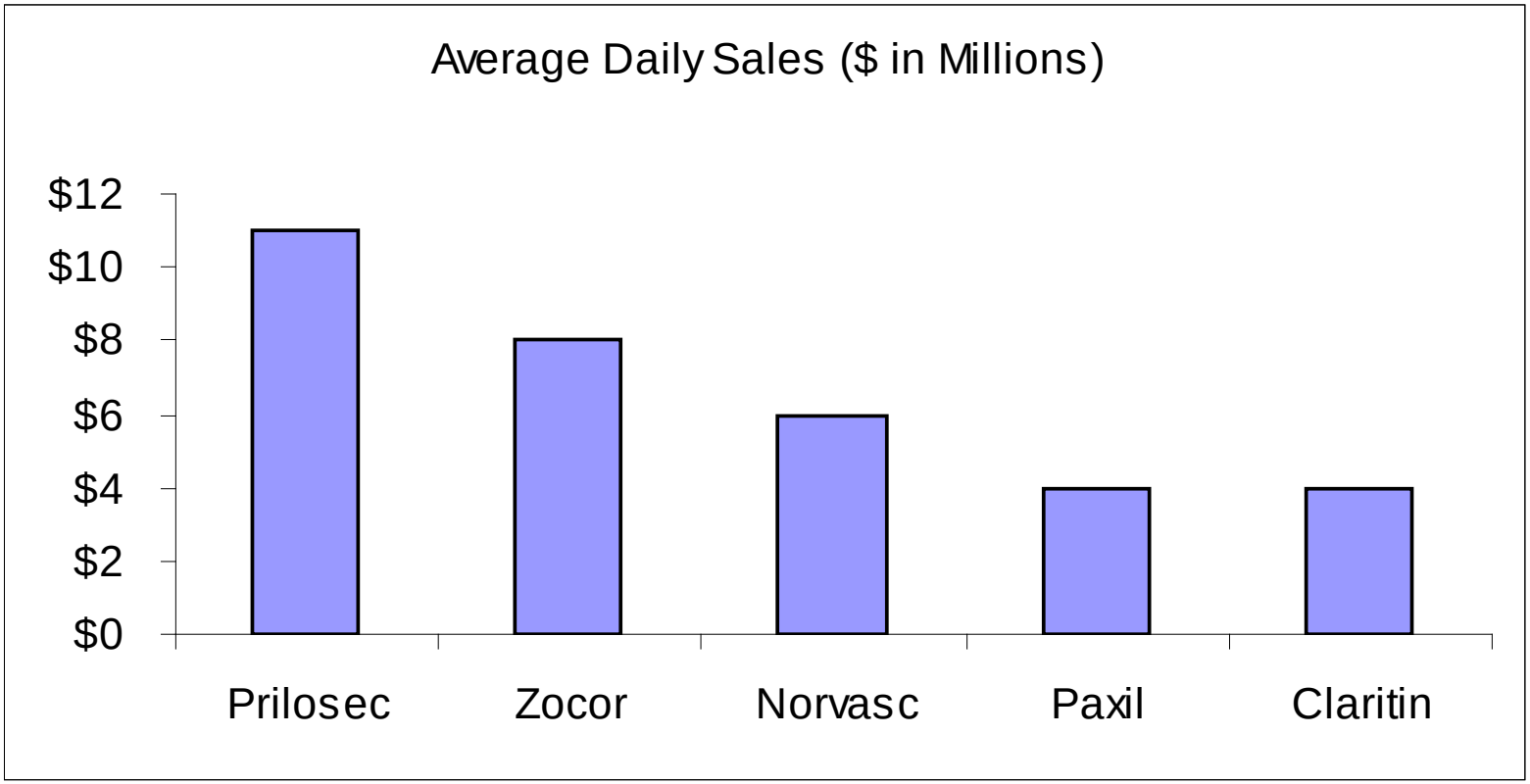
Wcześniejsze przerywanie próby

- The Helsinki Declaration states:

  "Physician should cease any investigation if the hazards are found to outweigh the potential benefits."

- Trials with serious, irreversible endpoints should be stopped if one treatment is "proven" to be superior
  - Possibility for early stopping formally included in the trial design.
  - Necessity of early stopping assessed by a « *Data and Safety Monitoring Board* » or « *Independent Data Monitoring Committee* ».

Koszta wydłużania procesu opracowywania leku



Próby z ustaloną liczebnością próbki…

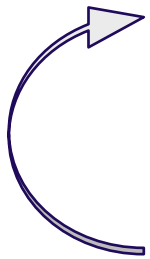
1 – Liczebność umożliwiająca „wykrycie” z ustaloną mocą określonej różnicy w skuteczności leczenia dla konkretnego poziomu istotności

2 – Wymagana liczba chorych włączana do próby

3 – Wyniki leczenia chorych analizowane na koniec próby, po zaobserwowaniu zaplanowanej (wymaganej) liczby zdarzeń

…vs próby z sekwencyjnymi grupami…

1 – Liczebność umożliwiająca „wykrycie” z ustaloną mocą określonej różnicy w skuteczności leczenia dla konkretnego poziomu istotności



2 – Chorzy włączani do próby do chwili analizy przejściowej

3a – Próba jest przerywana lub

3b – kontynuowana bez zmian

4 – Wyniki leczenia chorych analizowane na koniec próby, po zaobserwowaniu zaplanowanej (wymaganej) liczby zdarzeń

…vs próby adaptacyjne

1 – Liczebność umożliwiająca „wykrycie” z ustaloną mocą określonej różnicy w skuteczności leczenia dla konkretnego poziomu istotności
2 – Chorzy włączani do próby do chwili analizy przejściowej
3a – Próba jest przerywana lub
3b – jest kontynuowana bez zmian lub
3c – jest kontynuowana z modyfikacjami
4 – Wyniki leczenia chorych analizowane na koniec próby, po zaobserwowaniu zaplanowanej lub zmodyfikowanej liczby zdarzeń

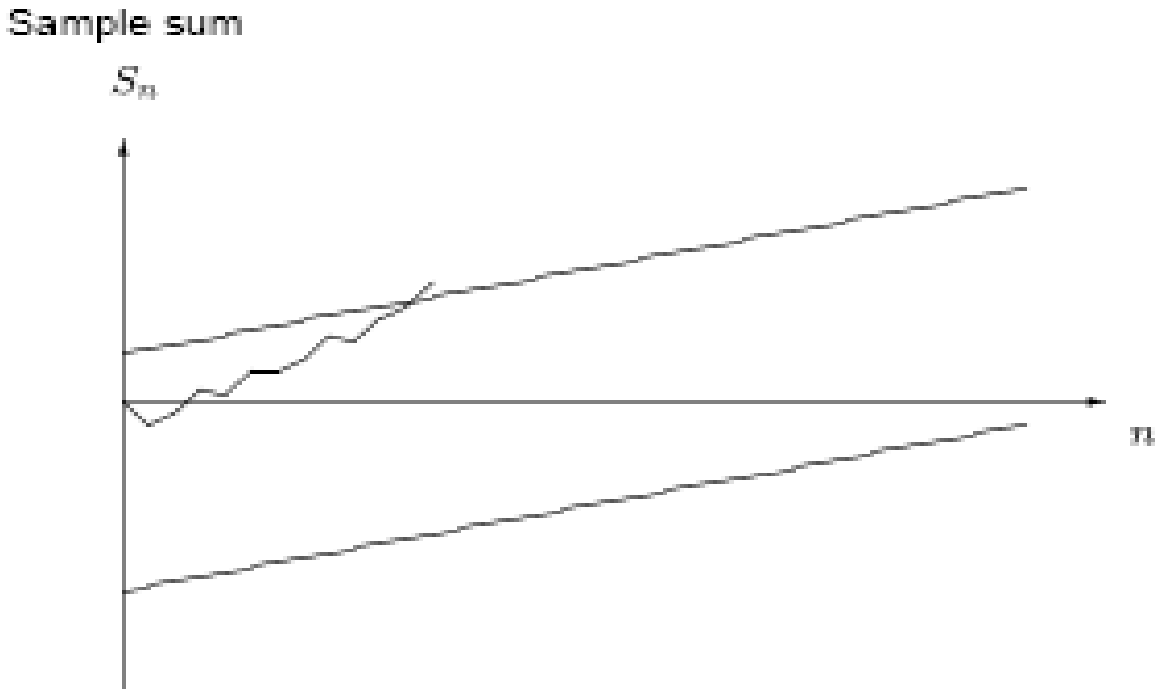
Próby sekwencyjne

- Analiza danych przeprowadzana po uzyskaniu każdej nowej obserwacji (konieczne ciągłe monitorowanie zdarzeń; rzadko możliwe)
- Granica decyzyjna (boundary) wyznaczana w oparciu o Z (stattystykę testową) i V (wariancję statystyki, rosnącą w czasie)

Whitehead, *The Design and Analysis of Sequential Clinical Trials* (1983)

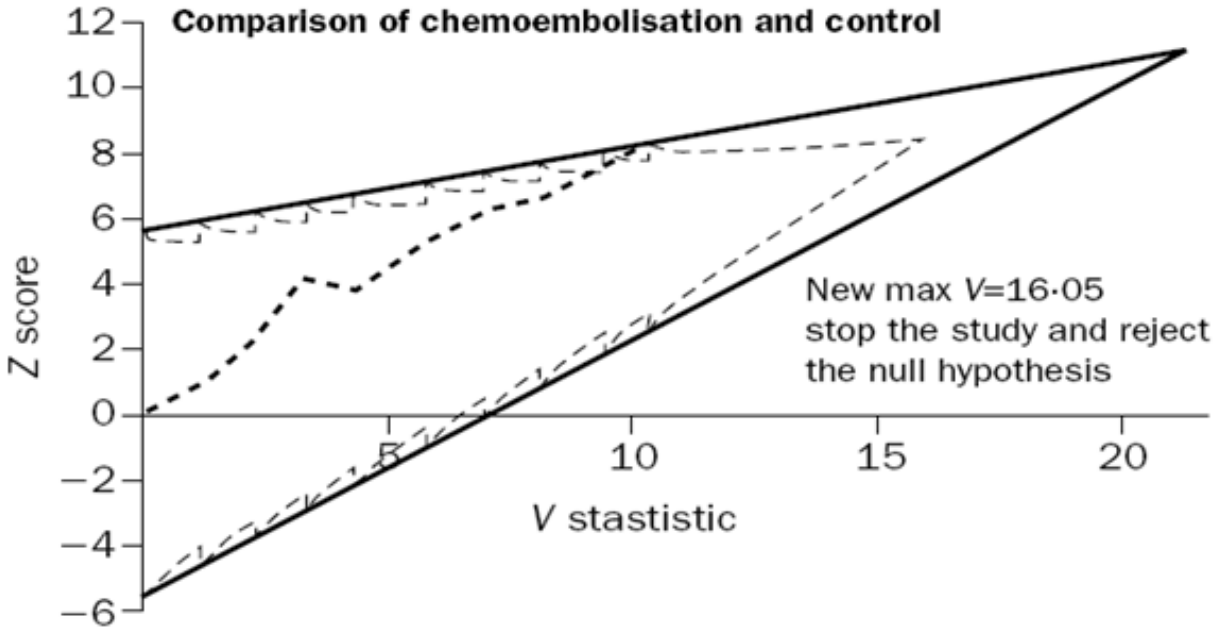
Próby sekwencyjne

- Wald (1947): Sequential Probability Ratio Test
  - Statystyką jest suma obserwacji
  - Bez górnego ograniczenia liczebności próbki



Próby sekwencyjne

- Whitehead (1983): Triangular Test
  - Example: effect of chemoembolization on 2-year survival in unresectable hepatocarcinoma

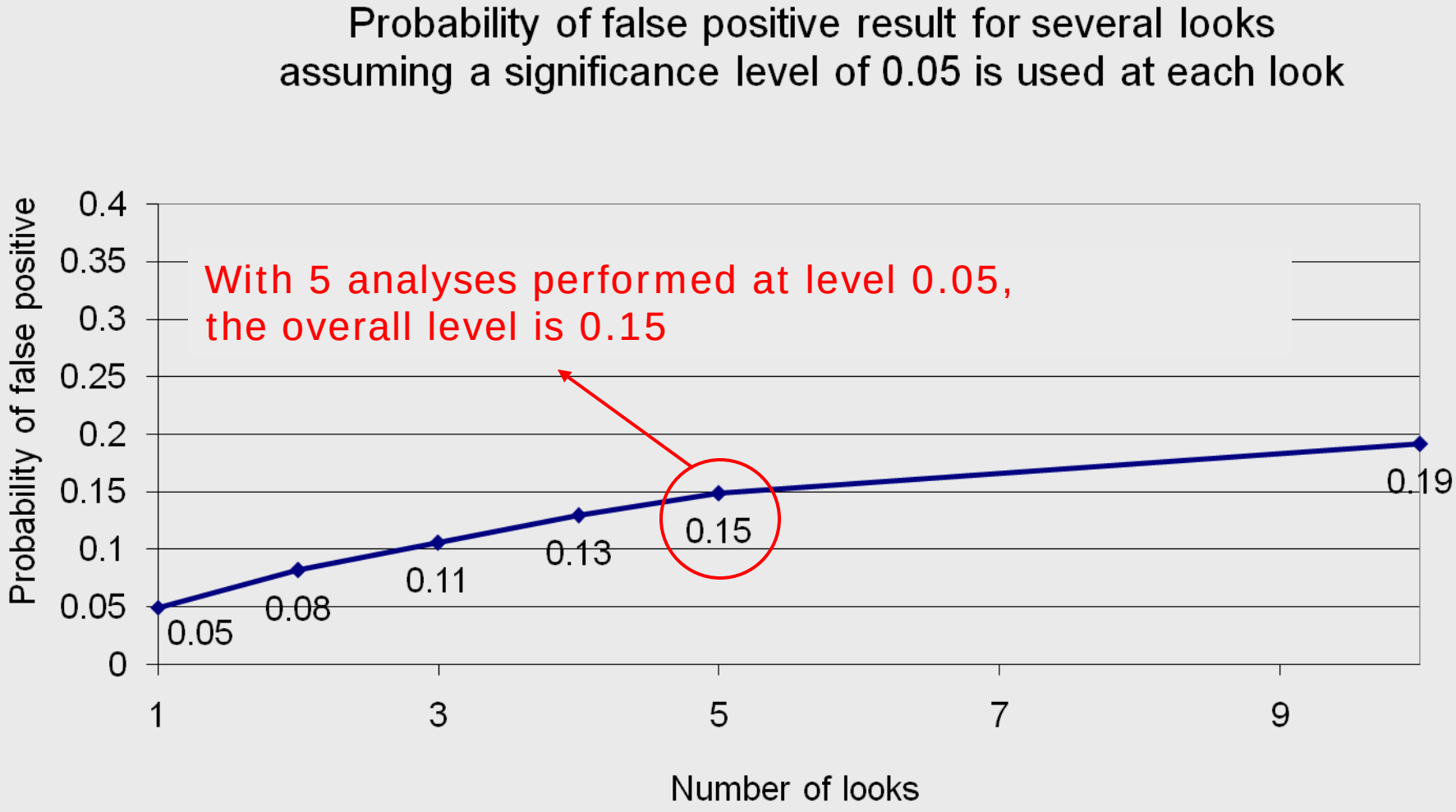


Llovet, *Lancet* (2002)

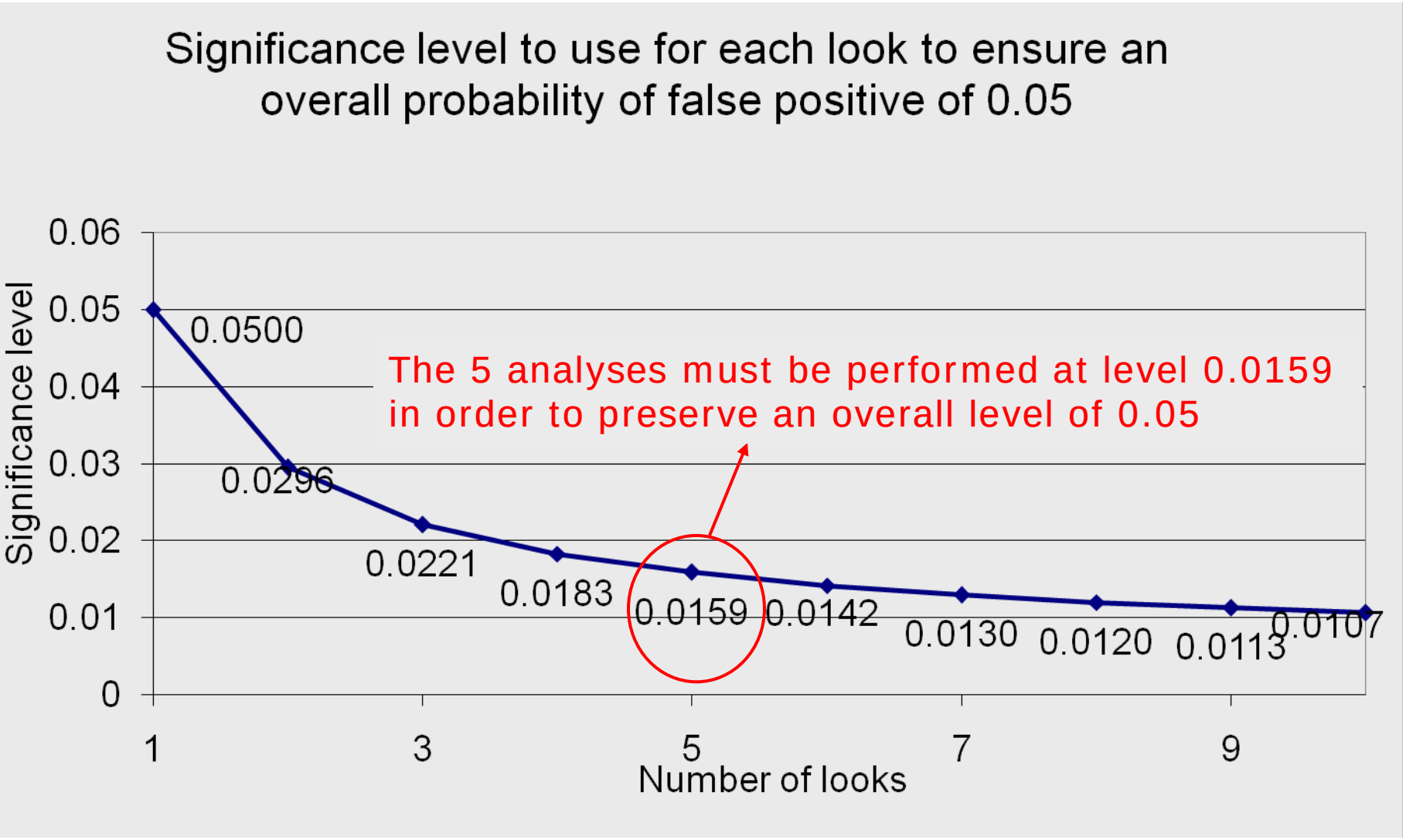
Próby z sekwencyjnymi grupami

- W praktyce, analizy danych z próby klinicznej możliwe są tylko co jakiś czas.
- Nawet dla tylko dwóch analiz, p-stwo błędu I rodzaju rośnie jeśli w analizach używany jest ten sam (nominalny) poziom istotności.
- W analizach przejściowych konieczne jest więc używanie skorygowanego poziomu istotności w celu kontroli całkowitego p-stwa błędu I rodzaju na ustalonym poziomie.

Inflation of α with multiple analyses



Adjusting α for multiple analyses



Układy doświadczalne z sekwencyjnymi grupami

- Testujemy $H_0: \Delta = 0$ vs. $H_A: \Delta \neq 0$
- m chorych włączanych do każdego z dwóch ramion próby pomiędzy analizami przejściowymi
- Rozważamy standaryzowane statystyki testowe Z_k, $k=1,...,K$

$$Z_k = \frac{\sum_{i=1}^{mk} X_{Ei} - \sum_{i=1}^{mk} X_{Ci}}{\sigma\sqrt{2mk}} = \frac{\bar{X}_{Ek} - \bar{X}_{Ck}}{\sigma\sqrt{2k/m}}$$

Układy doświadczalne z sekwencyjnymi grupami: błąd I rodzaju

- P-stwo błędnego przerwania próby/odrzucenia H_0 w *k*-tej analizie

  $P_{H0}(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k) = \pi_k$

  - „Błąd I rodzaju wydatkowany w *k*-tej analizie”

- $P(Błąd\ I\ rodzaju) = \sum \pi_k$

- Wybieramy c_k tak, aby $\sum \pi_k = \alpha$

Układy doświadczalne z sekwencyjnymi grupami: błąd II rodzaju

- P-stwo błędu II rodzaju wynosi

$$1\text{-}P_{HA}(U \{|Z_1|<c_1, \ldots, |Z_{k-1}|<c_{k-1}, | Z_k |\geq c_k\})$$

- Zależy od K, α, β, c_k.

- Ustalając te parametry, można wyznaczyć wymaganą liczebność próbki
  - da się ja wyrazić jako *R x (fixed sample size)*

Łączny rozkład p-stwa oszacowań miar efektu leczenia

- Załóżmy, że interesuje nas miara Δ.
- Niech $\hat{\Delta}_k$ będzie oszacowaniem w *k*-tej analizie.
- Informacja o Δ w *k*-tej analizie wynosi

$$I_k = 1/\operatorname{Var}\left(\hat{\Delta}_k\right)$$

- Dla różnych typów kryteriów oceny skuteczności leczenia, łączny rozkład p-stw oszacowań Δ jest w przybliżeniu wielowymiarowym rozkładem normalnym.

Łączny rozkład p-stwa standaryzowanych statystyk testowych

- Rozważamy test H_0: $\Delta = 0$ w k-tej analizie przy użyciu standaryzowanej statystyki testowej Z_k:

$$Z_k = \hat{\Delta}_k / \sqrt{\mathrm{Var}(\hat{\Delta}_k)} = \hat{\Delta}_k \sqrt{I_k}$$

- $(Z_1, \ldots, Z_K)$ ma w przybliżeniu wielowymiarowy rozkład normalny:

$$Z_k \sim \mathrm{N}(\Delta\sqrt{I_k}, 1)$$

$$\mathrm{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{I_{k_1} / I_{k_2}} \quad \text{for} \quad k_1 < k_2$$

Łączny rozkład p-stwa statystyk testowych „score”

- Rozważamy statstyki „score” $S_k = Z_k\sqrt{I_k}$:

$$S_k \sim \mathrm{N}(\Delta I_k, I_k)$$

- Mają własność „niezależnych przyrostów”:

$$\mathrm{Cov}(\mathrm{S}_k - \mathrm{S}_{k-1}, \mathrm{S}_{k'} - \mathrm{S}_{k'-1}) = 0 \text{ for } k \neq k'$$

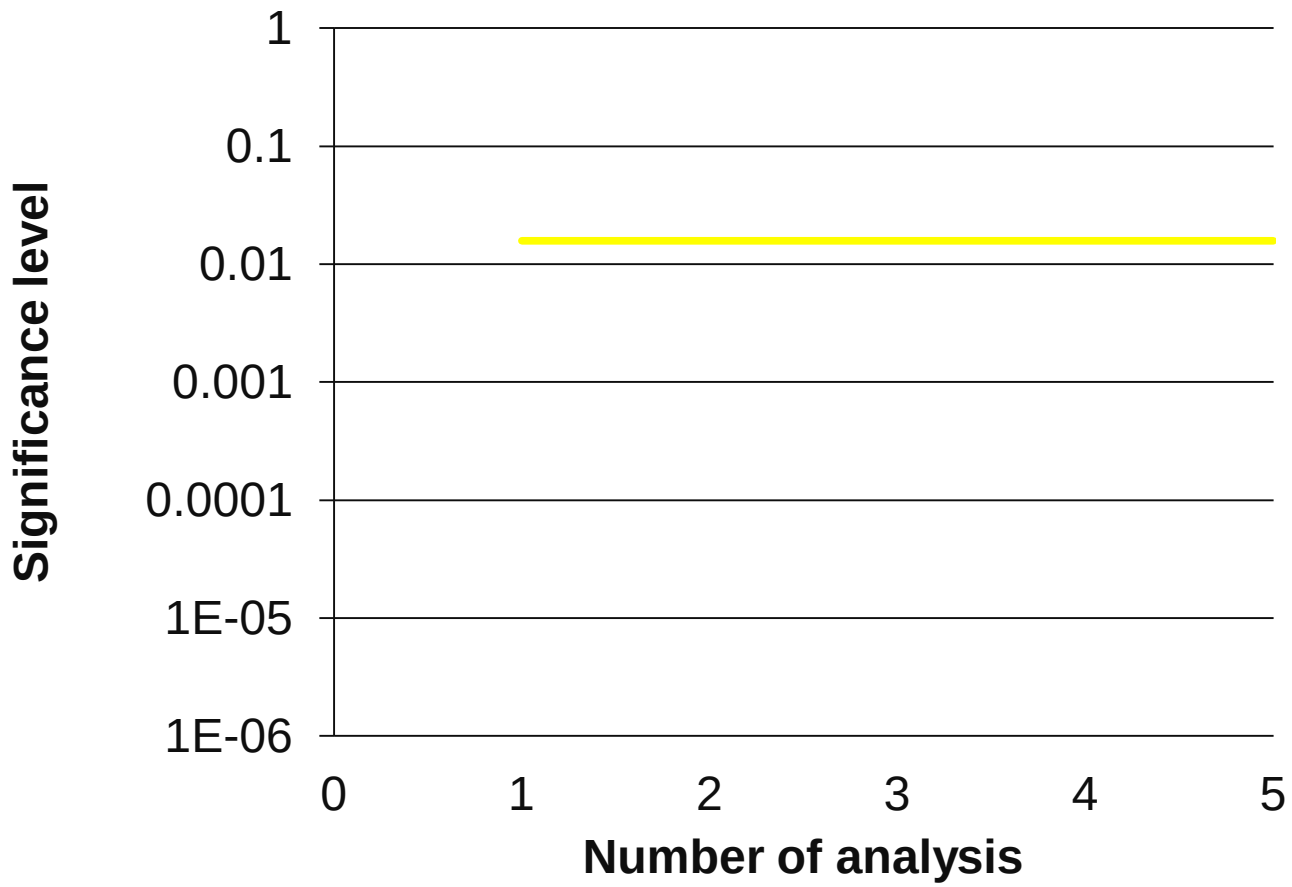
- Zachowuje się również dla Z_k …
- … i dla różnych kryteriów oceny skuteczności leczenia (ciągłych, binarnych, czasu do zdarzenia,…)
- Umożliwia obliczanie p-stwa błędu

Granice decyzyjne Pococka

- Odrzucamy H_0 jeśli $|Z_k| > c_P(K,\alpha)$
  - $c_P(K,\alpha)$ wybierane tak, aby P(Błąd I rodzaju) = α
- Wszystkie analizy przeprowadzane dla tego samego, skorygowanego poziomu istotności
- Relatywnie duże p-stwo wcześniejszego przerwania próby, ale moc analizy końcowej może być zmniejszona

Granice decyzyjne Pococka

- Signif. levels for Z_k (2-sided) per interim analysis (K=5)

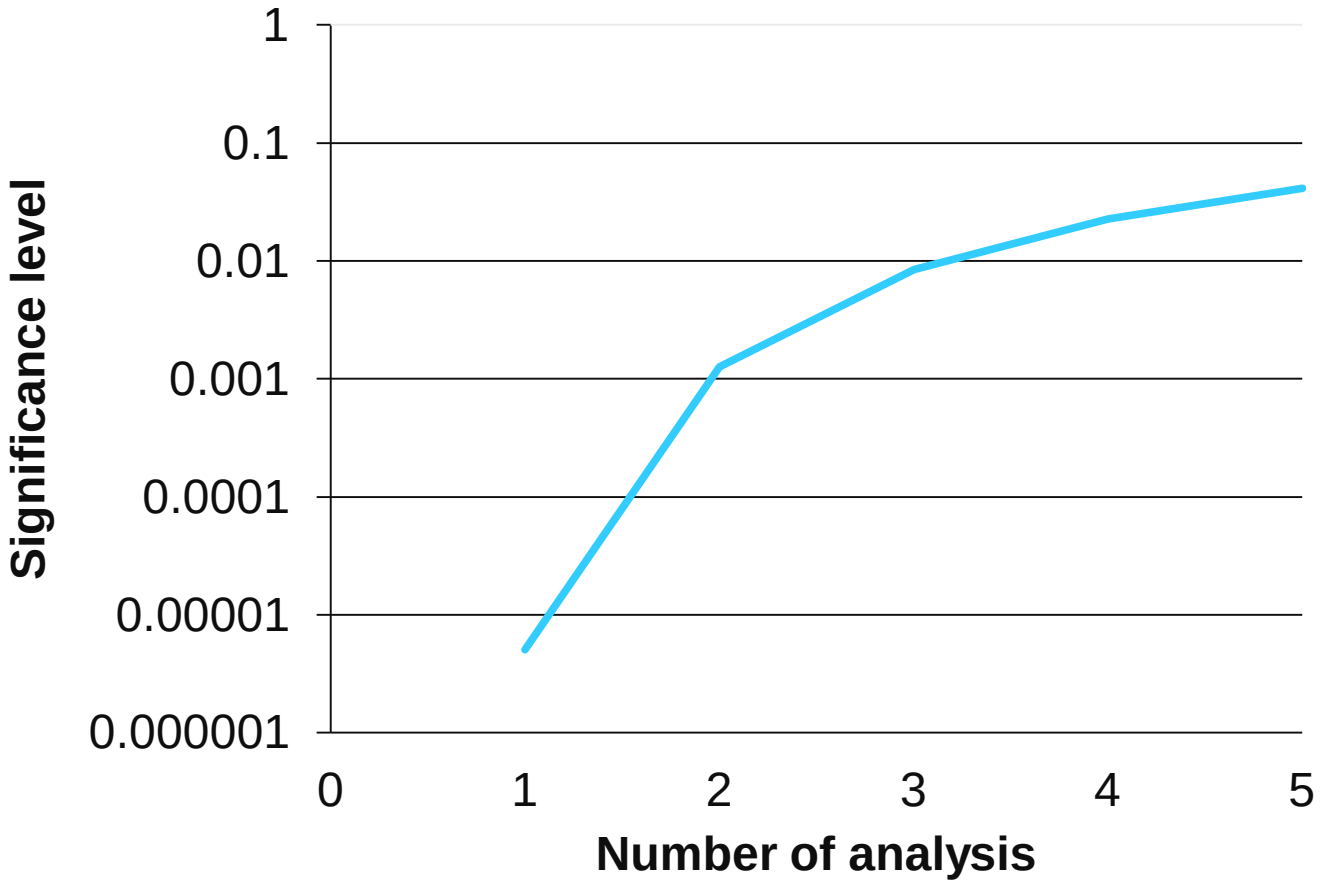


Granice decyzyjne O'Briena-Fleminga

- Odrzucamy H_0 jeśli $|Z_k| > c_{OBF}(K,\alpha)\sqrt{(K/k)}$
  - Dla $k=K$ mamy $|Z_K| > c_{OBF}(K,\alpha)$
  - $c_{OBF}(K,\alpha)$ wybierane tak, aby $P(\text{Błąd I rodzaju}) = \alpha$
- „Wczesne" analizy przy użyciu mocno skorygowanego poziomu istotności
- Relatywnie małe p-stwo przerwania próby, ale moc analizy końcowej praktycznie niezmieniona

Granice decyzyjne O'Briena-Fleminga

- Signif. levels for Z_k (2-sided) per interim analysis (K=5)



Granice decyzyjne Wanga-Tsiatisa

- Wang & Tsiatis (1987):

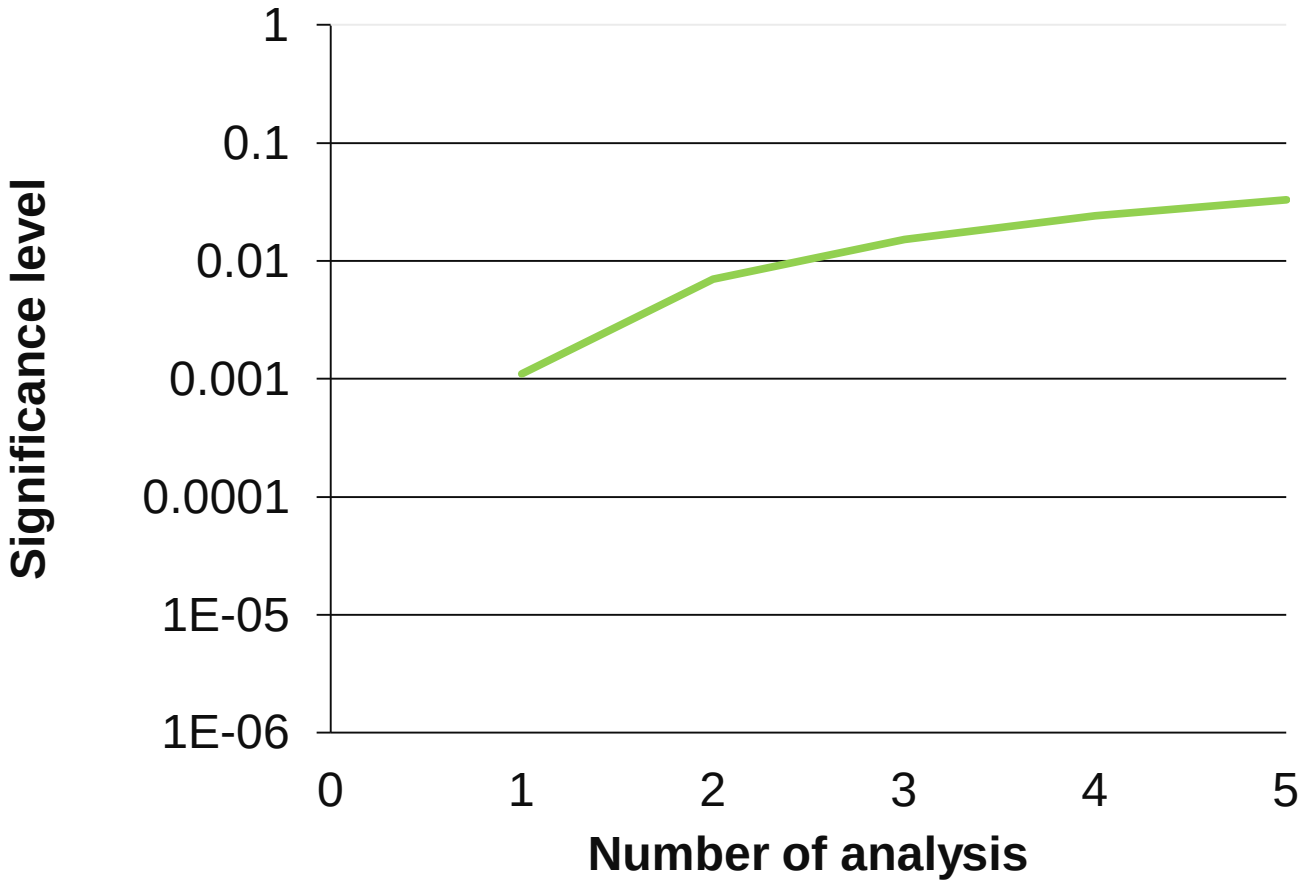
  Odrzucamy H_0 jeśli $| Z_k | > c_{WT}(K,\alpha,\theta)(k / K)^{\theta - ½}$

  - $\theta = 0.5$ daje granice Pococka; $\theta = 0$, O'Briena-Fleminga
  - dostępne w EaSt

- Umożliwiają wybór rozwiązań pośrednich pomiędzy granicami Pococka i O'B-F

Granice decyzyjne Wanga-Tsiatisa

- Signif. levels for Z_k (2-sided) per interim analysis (K=5) with $\theta = 0.2$



Granice decyzyjne Haybittle-Peto

- Haybittle & Peto (1976):

  Odrzucamy H_0 jeśli $|Z_k| > 3$ dla $k = 1,...,K\text{-}1$

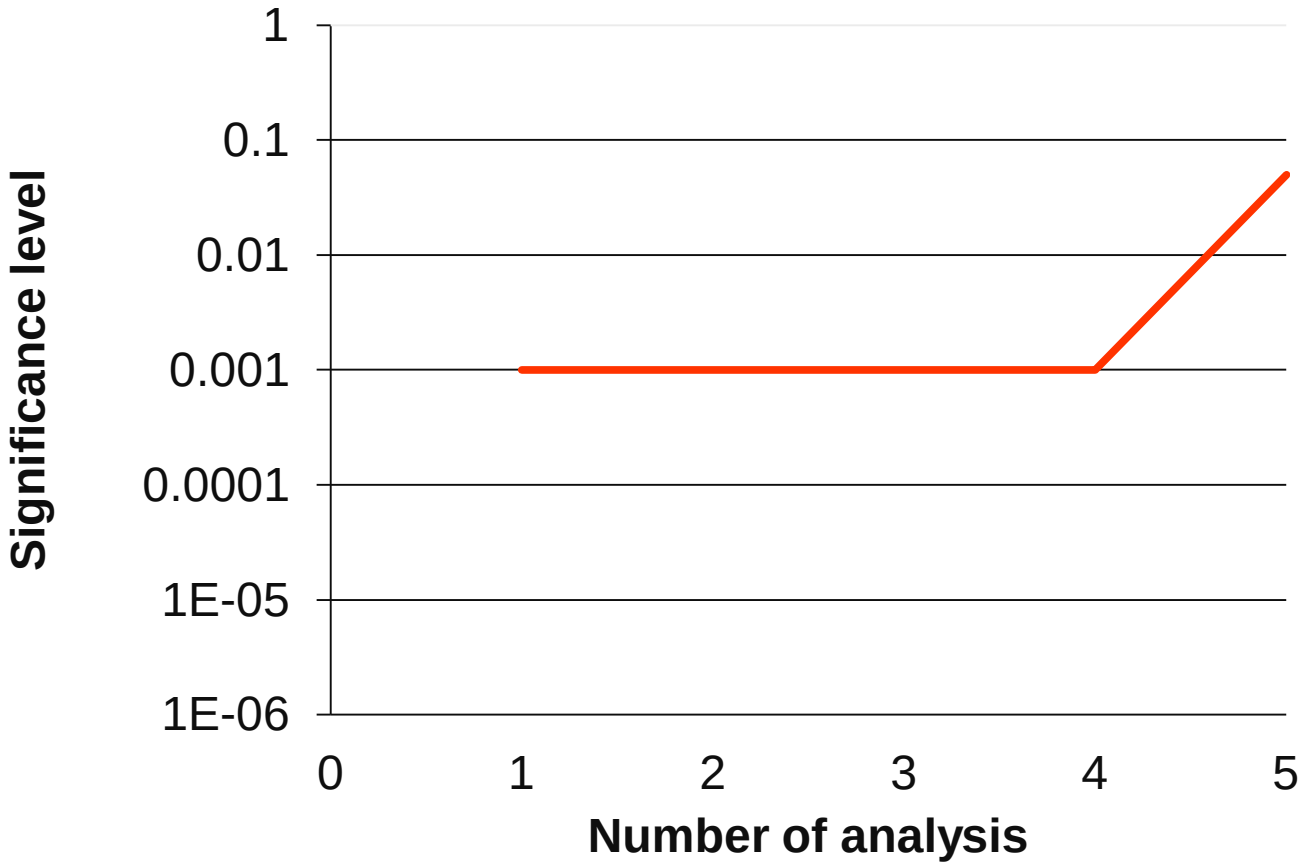
  Odrzucamy H_0 jeśli $|Z_k| > c_{HP}(K,\alpha)$ dla $k = K$

  - $|Z_k| > 3$ odpowiada użyciu $p < 0.0026$

- „Wczesne” analizy dla mocno obniżonego, ale akceptowalnego poziomu istotności
- Intuicyjne podejście, łatwe do zaimplementowania (pomijając korektę dla analizy końcowej)

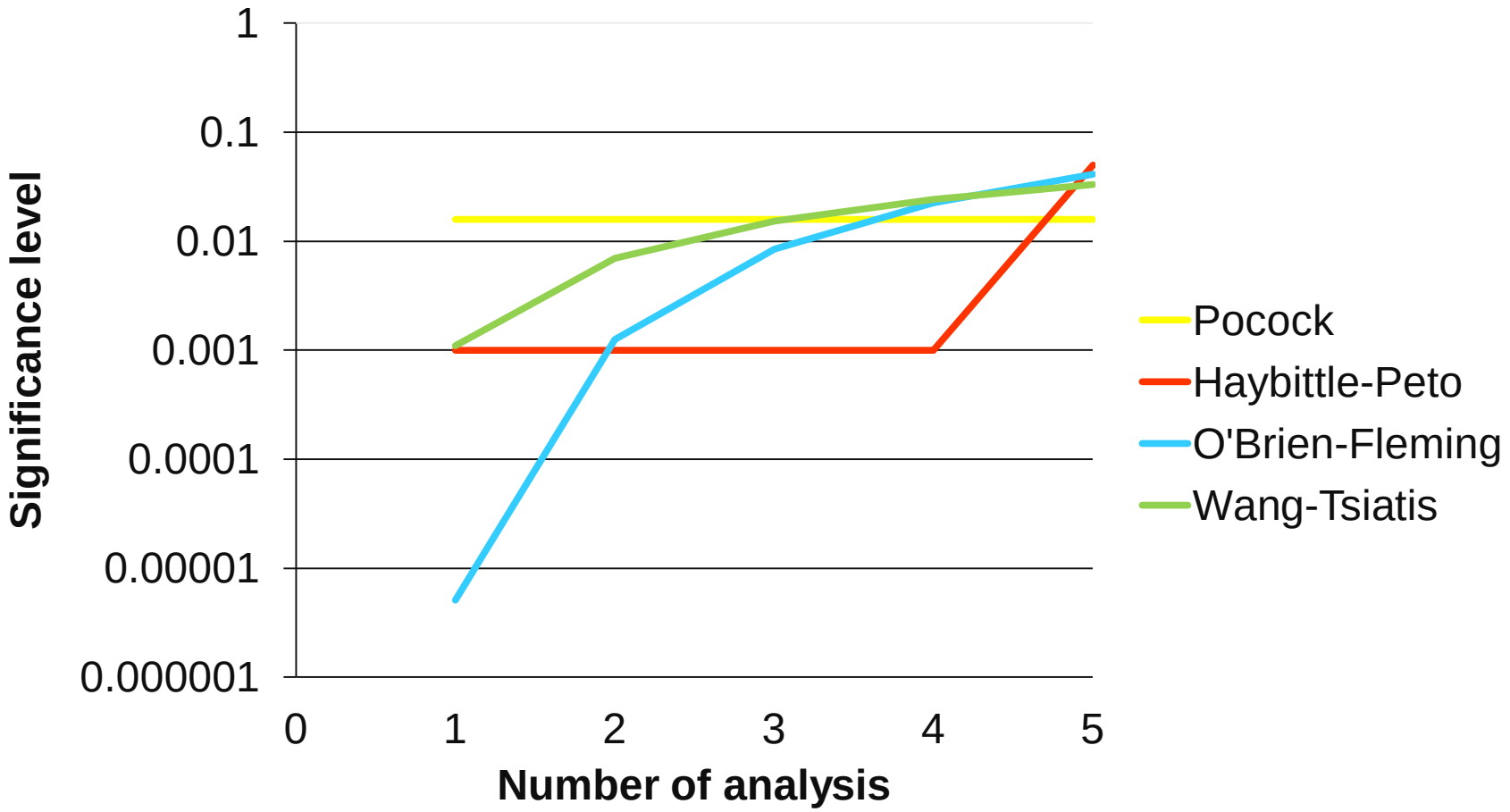
Granice decyzyjne Haybittle-Peto

- Signif. levels for Z_k (2-sided) per interim analysis (K=5)



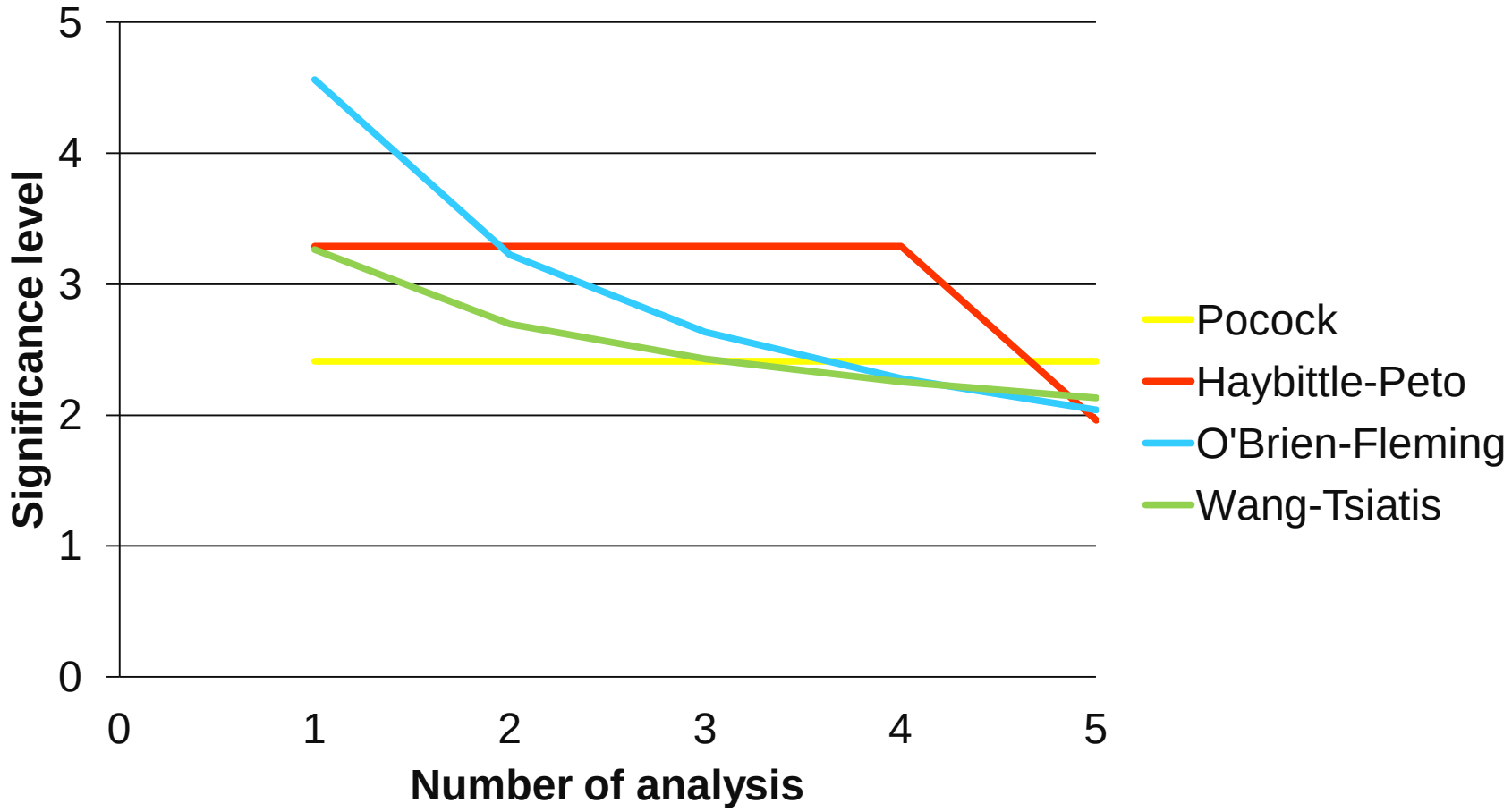
Porównanie różnych granic decyzyjnych

- Signif. levels for Z_k (2-sided) per interim analysis (K=5)



Porównanie różnych granic decyzyjnych

- Z_k per interim analysis (K=5)



Potencjalne oszczędności / straty dla użycia prób z sekwencyjnymi grupami

Oczekiwane liczebności próbki dla $K=5$:

- kryterium o rozkładzie normalnym z $\sigma = 2$
- $\alpha = 0.05$
- $\beta = 0.1$ dla $|\mu_A - \mu_B| = 1$

$\lvert \mu_A - \mu_B \rvert$	Fixed sample	Pocock	O'Brien-Fleming
0.0	170	205	179
0.5	170	182	168
1.0	170	117	130
1.5	170	70	94

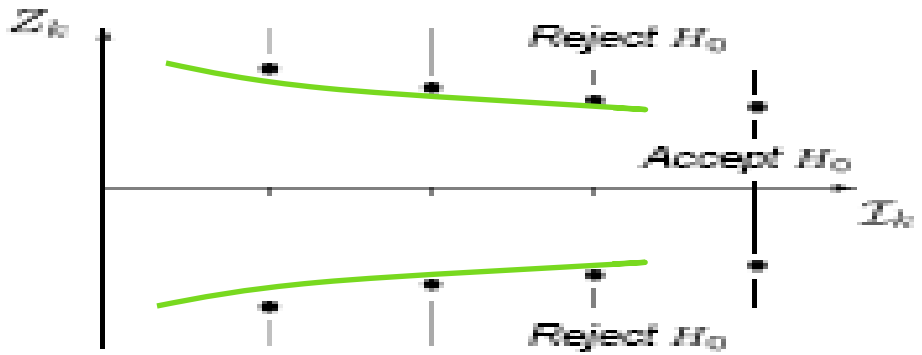
Wcześniejsze przerywanie próby

- W celu **odrzucenia** H_0 o *braku efektu leczenia*
  - Uniknięcie podawania mniej skutecznego leczenia kolejnym chorym
  - Sensowne jeśli nie trzeba gromadzić dodatkowych danych o, np., toksyczności czy efektach długoterminowych.

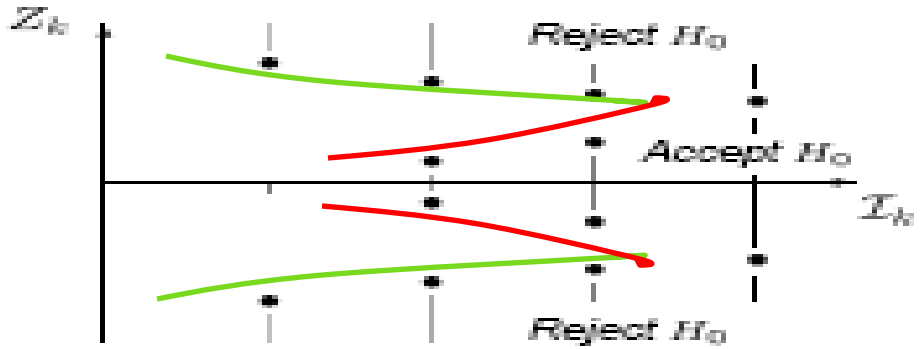
- W celu **przyjęcia** H_0 o *braku efektu leczenia*
  - Stopping “for futility” or “abandoning a lost cause”
  - Oszczędza czas i środki gdy próba z dużym p-stwem nie przyniesie pożądanych wyników.

Test dwustronny

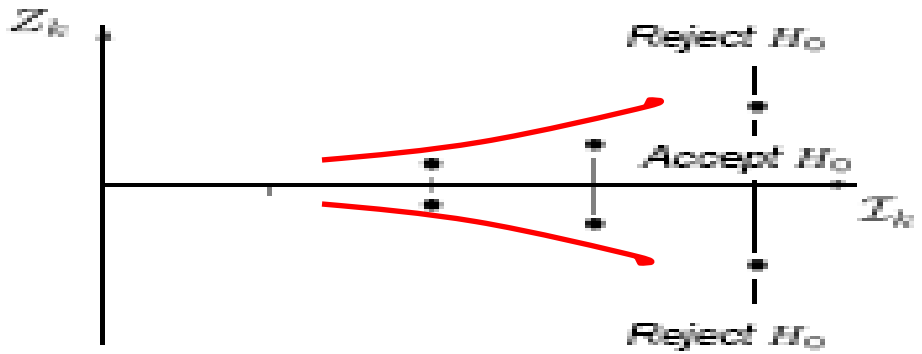
Early stopping to reject H_0



An inner wedge:
Early stopping to reject H_0 or accept H_0

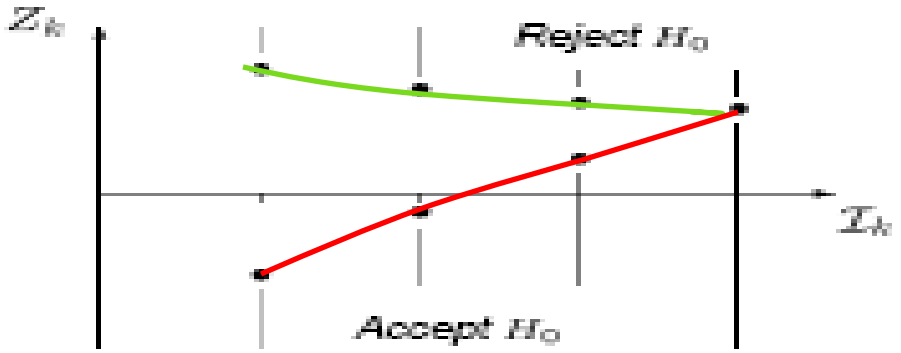


Abandoning a lost cause:
Only an inner wedge

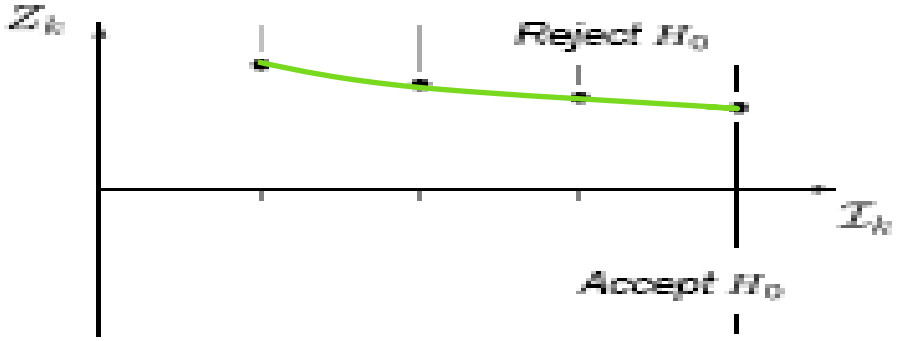


Test jednostronny

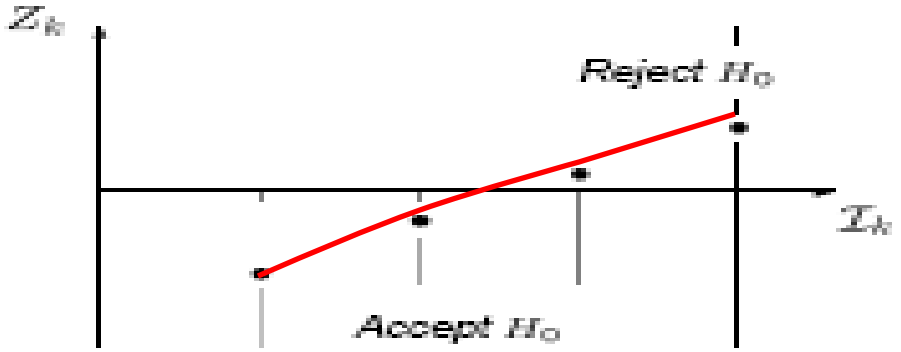
Early stopping to reject H_0 or accept H_0



Early stopping only to reject H_0



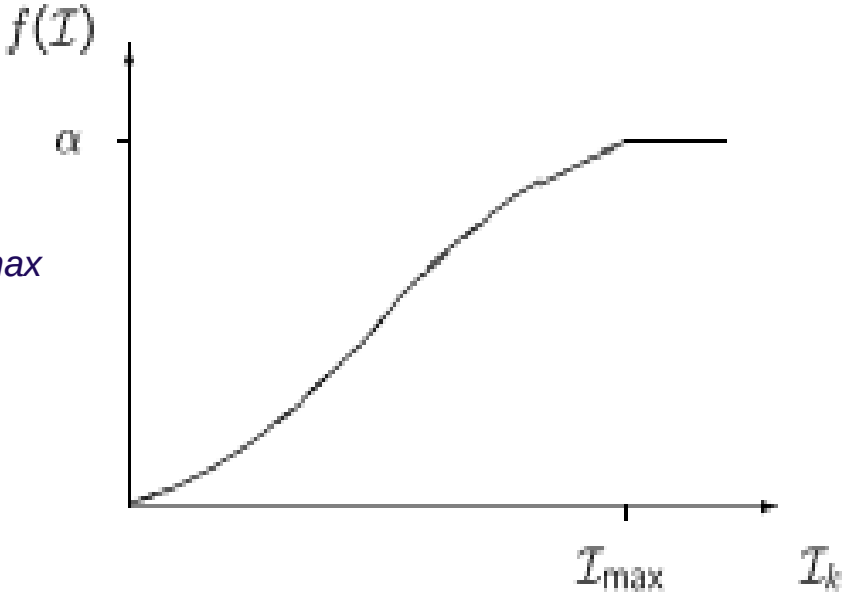
Abandoning a lost cause:
Early stopping only to accept H_0



Strategia „wydatkowania błędu” (error-spending)

- Usuwa wymóg ustalania liczby analiz w równych odstępach
- Lan & DeMets (1983): „wydatkowanie” błędu I rodzaju

- *Układ dośw. z maksymalną informacją:*
  - Funkcja wydatkowania błędu →
  - Określa granice decyzyjne
  - Akceptujemy H_0 jeśli osiągamy I_{max} bez odrzucenia hipotezy zerowej

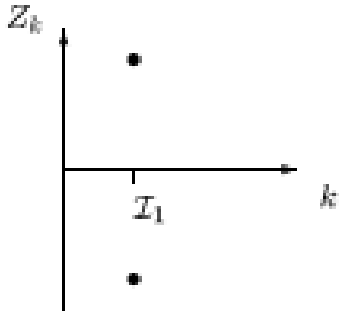


Analysis 1:

Observed information $\mathcal{I}_1$.

Reject H_0 if $|Z_1| > c_1$ where

$$Pr_{\theta=0}\{|Z_1| > c_1\} = f(\mathcal{I}_1).$$

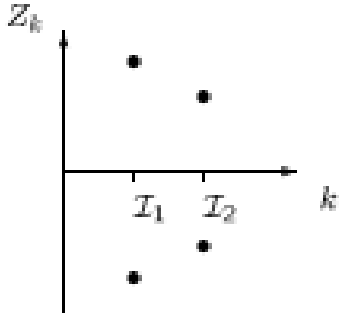


Analysis 2:

Cumulative information $\mathcal{I}_2$.

Reject H_0 if $|Z_2| > c_2$ where

$$Pr_{\theta=0}\{|Z_1| < c_1, |Z_2| > c_2\}$$
$$= f(\mathcal{I}_2) - f(\mathcal{I}_1).$$



etc.

Strategia ”wydatkowania błędu”

- $f(t)=min(2-2\Phi(z_{1-\alpha/2}),\alpha)$ daje ≈ granice O’B-F
- $f(t)=min(\alpha\ ln\ (1+(e\ -1)t,\alpha)$ daje ≈ granice Pococka
- $f(t)=min(\alpha t^{\theta},\alpha)$:
  - $\theta = 1$ odpowiada granicom Pococka, 3 granicom O’B-F

Ile analiz przejściowych?

- Jedna lub dwie dają największą część zysku w terminach redukcji oczekiwanej liczebności próbki
- Minimalny zysk dla więcej niż 5 analiz

Kiedy przeprowadzać analizy?

- Używając strategii „wydatkowania błędu”, pełna elastyczność co do liczby i terminu analiz
  - pierwsza nie powinna być „za wcześnie” (na ogół gdy mamy ~ 50% całkowitej informacji)
  - analizy w rónych odstępach zalecane z przyczyn praktycznych
- Strategia/termin analiz nie powinny być wybierane na podstawie zaobserwowanych wyników

Analiza końcowa w próbie z sekwencyjnymi grupami

- Dla testów dwustronnych i wczesnego przerywania na korzyść H_A, estymator najw. wiarogodności na ogół przeszacowuje Δ
  - dodatnie obciążenie dla $\Delta > 0$, ujemne dla $\Delta < 0$
  - bo przerywamy próbę jeśli oszacowanie jest duże
  - opracowano estymatory korygujące obciążenie

Independent Data Monitoring Committee (IDMC)

- Niezależny od organizatorów próby
- Eksperci z różnych dyscyplin
  - klinicyści, statystycy, etycy, …
- Ochrona interesów i bezpieczeństwa chorych, przy jednoczesnym zapewnieniu naukowej wiarygodności próby

IDMC: zadania

- Ocena porównywalności ramion próby
- Monitorowanie tempa rekrutacji i czasu trwania próby
- Monitorowanie jakości gromadzonych danych
- Monitorowanie bezpieczeństwa/toksyczności
  - Na ogół bez „zaślepiania”
- Ocena różnic w skuteczności leczenia
  - „Zaślepiona” całkowicie lub częściowo (np. przez użycie kodów X /Y dla porównywanych metod leczenia)

IDMC: pytania

- Czy próba powinna być kontynuowana?
  - bezpieczeństwo
  - skuteczność
  - wyniki innych prób
- Czy protokół powinien zostać zmodyfikowany?

Modyfikacja liczebnośći próki

- Dla kryteriów oceny skuteczności o rozkładzie normalnym

$$n_I = \frac{2\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2}{\left(\frac{\Delta}{\sigma}\right)^2}$$

  - Liczebność próbki zależy od σ^2
  - Dla błędnej wartości, n_I może być zbyt małe
  - Idea: „wewnętrzne badanie pilotażowe”
    - szacujemy σ^2 z danych uzyskanych na początku próby
    - wyznaczamy nową luiczebność póbki, n_A
    - jeśli konieczne, włączamy więcej niż n_I chorych

Wewnętrzne badanie pilotażowe

- Szacujemy σ^2 przez s^2_0 z początkowych danych
  - Wittes & Brian (1990): „tradycyjne” oszacowanie (bez zaślepiania)
  - Kieser & Friede (2003): oszacowanie zaślepione

- Wyznaczamy nową liczebność próbki, n_A
  - Wittes & Brian (1990): przyjmujemy $n = max(n_I, n_A)$
  - Birkett & Day (1994): $n = max(aktualne\ n, n_A)$
  - Gould & Shih (1992): $n = min(n_A, 2n_I)\ jeśli\ n_A > 1.25n_I$

Wewnętrzne badanie pilotażowe

- Dla małych badań pilotażowych i n_I, p-stwo błędu I rodzaju wzrasta dla modyfikacji *bez zaślepienia*
  - przynajmniej 10 chorych/grupa w badaniu pilotażowym
- Dla *zaślepionej* modyfikacji, generalnie nie ma problemu
- Oczekiwana moc nieco mniejsza dla obu metod

- Modyfikacja nie powinna opierać się na oszacowaniach miary efektu leczenia
  - tylko na oszacowaniach wariancji

Binarne kryteria oceny skuteczności leczenia

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 2\bar{\pi}\left(1-\bar{\pi}\right)}{\Delta^2}$$

$$\bar{\pi} = \frac{\pi_E + \pi_C}{2}$$

- Gould & Shih (1992): oszacowanie pilotażowe π na połączonych danych z obu grup
  - zaślepione
- Herson & Wittes (1993): szacujemy π_C z danych pilotażowych, a π_E przez (oszacowane π_C) + Δ
  - bez zaślepiania

Stochastyczne „obcinanie”

- Przerywamy próbę, jeśli p-stwo odrzucenia hipotezy zerowej, warunkowo ze względu na zaobserwowane dane, jest wysokie
  - mocy warunkowa
  - moc przewidywana
  - podejście nieparametryczne

Moc warunkowa

- Rozważamy „test odniesienia” T
  - dla testowania hipotezy zerowej H_0: $\Delta=0$
- Dla *k*-tej analizy, definiujemy

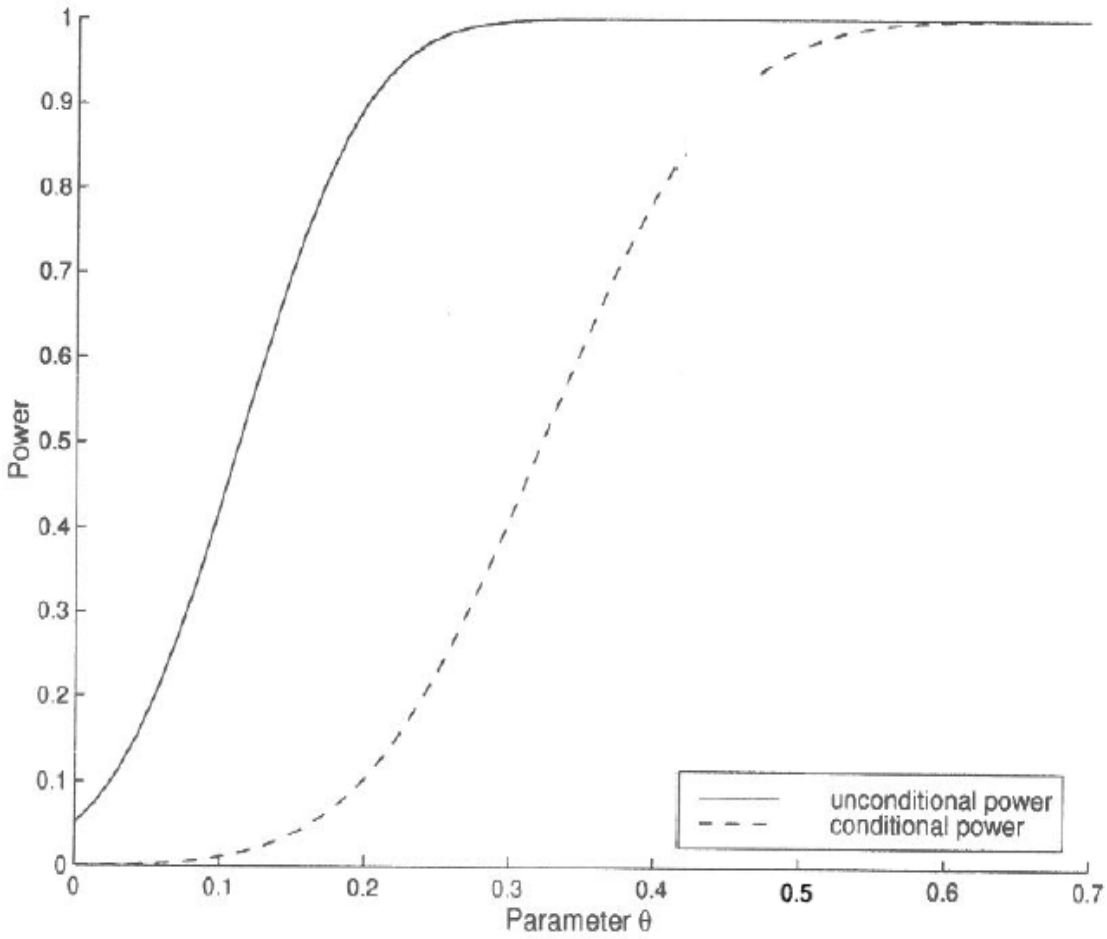
  $p_k(\Delta) = P_{HA}(\text{test odrzuci } H_0 \mid \text{zgromadzone dane})$
- Duża wartość $p_k(0)$ sugeruje, że T odrzuci H_0
  - przerywamy próbę, odrzucamy H_0 dla $p_k(0) > \xi$=0.8 lub 0.9
  - przerywamy próbę, przyjmujemy H_0 dla $1-p_k(\Delta)>\xi'$ (1-sided)

    lub dla $1-p_k(-\Delta) > \xi'$ oraz $1-p_k(\Delta) > \xi'$ (2-sided)
- P-stwo błędu I rodzaju nie większe niż α / ξ
  - Dla błędu II rodzaju nie większe niż β / ξ'

Moc warunkowa

- Moc bezwarunkowa przy α=0.05 i β=0.1 dla Δ=0.2
- Moc warunkowa dla analizy przejściowej z oszacowaniem Δ=0.1
  - p-stwo odrzucenia h. zerowej na koniec próby zredukowane z 0.9 do 0.1

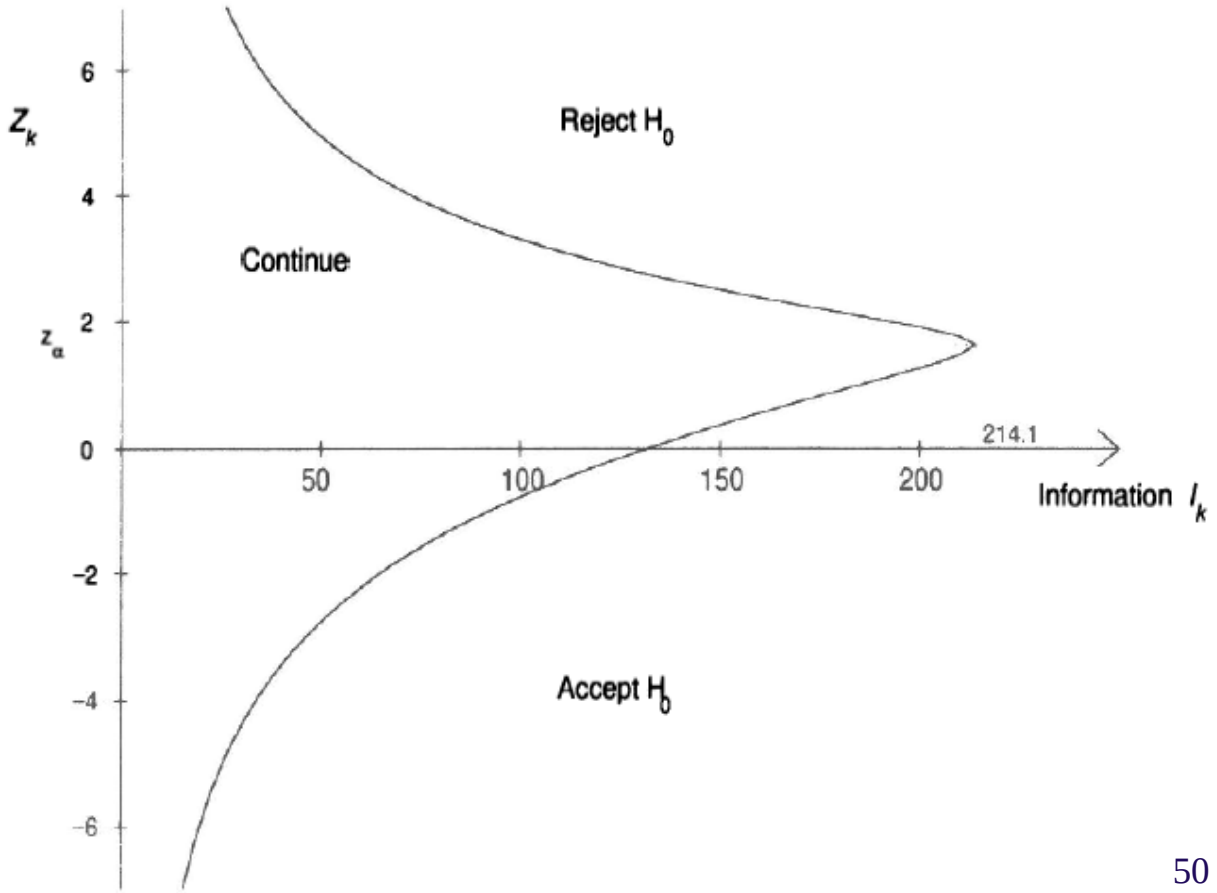
Figure 10.1 *Conditional and unconditional power curves for a one-sided test*



Moc warunkowa

- Granice decyzyjne odpowiadające stochastycznemu „obcinaniu”

Figure 10.2 *Stopping boundary for a stochastically curtailed one-sided test using th conditional power approach. The reference test is a fixed sample one-sided test with Type error probability* $\alpha = 0.05$ *and information level* $\mathcal{I}_{f,1} = 214.1$, *set to achieve power* 0.9 *a* $\theta = 0.2$. *The stochastic curtailment parameters are* $\gamma = \gamma' = 0.8$.



Moc warunkowa i przewidywana

- Problem z podejściem opartym na mocy warunkowej: obliczenia oparte na wartości Δ odbiegającej od aktualnego oszacowania.
- Rozwiązanie: uśrednienie po wartościach Δ
- „Moc przewidywana”

$$P_k = \int p_k(\Delta)\pi(\Delta \mid \text{data})d\Delta$$

  - $\pi(\Delta \mid \text{data})$ jest rozkładem *a posteriori*
  - Przerywamy próbę, odrzucamy H_0 dla $P_k > \xi$ itd.
  - Jaki rozkład *a priori* ?

Moc przewidywana

- Granice decyzyjne odpowiadające stochastycznemu „obcinaniu”
- Węższe niż dla mocy warunkowej
  - łatwiej przerwać próbę

Figure 10.4 *Stopping boundary for a stochastically curtailed one-sided test using the predictive power approach with a uniform prior. The reference test is a fixed sample one-sided test with Type I error probability* $\alpha = 0.05$ *and information level* $\mathcal{I}_{f,1} = 214.1$*, set to achieve power* 0.9 *at* $\theta = 0.2$*. The stochastic curtailment parameters are* $\gamma = \gamma' = 0.8$.