Ad 1.

Aby odpowiedzieć na pytanie, czy dwie próby (= grupy) pochodzą z tej samej populacji, porównamy ze sobą ich zmienność międzygrupową ($SS_B$) i wewnątrzgrupową ($SS_W$). $SS_B$ informuje o tym, jak znacznie różnią się między sobą średnie w dwóch grupach, zaś $SS_W$ jak duże jest zróżnicowanie wewnątrz samych grup. Przewaga na korzyść $SS_B$ świadczy o faktycznym różnieniu się między sobą grup.

Tożsame z takim porównaniem jest porównanie z sobą wariancji $s^2_B$ i $s^2_W$ - jako wielkości ściśle związanych z SS: **wariancja = SS / df** (df to liczba stopni swobody). Gdyby Czytelnik chciał sobie odświeżyć te elementarne fakty, proponuję sięgnąć do książki Grażyny Wieczorkowskiej "Statystyka – od teorii do praktyki" z Wydawnictwa Scholar (rozdział 5).

Porównanie z sobą wariancji jest możliwe dzięki statystyce Fishera (rozkład F). W statystyce niczego nie szacuje się "na oko", ani też "widać, że", tylko w oparciu o znane, matematycznie dowiedzione rozkłady danej statystyki.

Liczymy zatem **(do wszystkich obliczeń stosuję Zestaw I danych; kolumnie pierwszej odpowiada indeks 1, kolumnie drugiej – indeks 2)**:

$$M_1 = 149.9556;\ M_2 = 150.1620;\ M \text{ (z obu)} = 150.0588 .$$

$$SS_1 := \Sigma(x_{i,1} - M_1)^2 = 804.6219;\ SS_2 := \Sigma(x_{i,2} - M_2)^2 = 665.6668 .$$
$$SS_W := SS_1 + SS_2 = 1470.2980 .$$
$$df_W := \text{liczba grup} \cdot (\text{liczebność grupy} - 1) = 2 \cdot (35 - 1) = 68.$$
$$s^2_W = 21.6219 .$$

$$SS_B := (M_1 - M)^2 \cdot n_1 + (M_2 - M)^2 \cdot n_2 = 0.7457 .\ \text{(nb. liczebności n są takie same)}$$
$$df_B := \text{liczba grup} - 1 = 2 - 1 = 1 .$$
$$s^2_B = 0.7457 .$$

$$F = s^2_B / s^2_W = 0.0345 .$$

Z tablic odczytujemy, że obszar krytyczny F dla $\{df_B , df_W , \alpha = 0.05$ (poziom ufności 95%) $\}$ odcina wartość F = 3.98. Otrzymana przez nas wartość F = 0.0345 nie leży w obszarze krytycznym (jest mniejsza); zatem stwierdzamy **brak podstaw do odrzucenia (bpdo)** hipotezy zerowej o równości średnich w grupach. *Dla przypomnienia: hipoteza zerowa jest przeciwieństwem naszej hipotezy badawczej, tutaj: że dane w grupach się od siebie różnią. Używa się hipotezy zerowej, a nie badawczej, bowiem korzysta się z dowodu nie wprost – ad absurdum.* Nie mogąc odrzucić (na poziomie ufności 95%) hipotezy o równości średnich, uznajemy ich różnicę za **nieistotną statystycznie**, średnia 2 **nie jest istotnie statystycznie** większa od średniej 1.

Obliczenia te wolno nam było wykonać pod założeniem równości wariancji w podpopulacjach, co sprawdza się (za pomocą pakietu ststystycznego) testem Levene'a. Nie będziemy jednak badać słuszności tego założenia, ponieważ porównywanie ze sobą grup **równolicznych** jest odporne na niespełnienie założenia (wyjaśnienie tego faktu matematycznego wykracza zdecydowanie poza ramy zadania). Innymi słowy, do grup równolicznych zawsze można zastosować analizę rozkładu F, który (ze względu na treść obiektu badania – wariancję) nosi również nazwę analizy wariancji (ANOVA) (skądinąd jednoczynnikowej, o dwóch poziomach czynnika).

Ad 2.

Badamy losowość łącznej próby 1+2. Zastosujemy tu test losowości próby, oparty na liczbie serii (cf. te hasła w internecie, sporo wykładów e-learningowych). Hipoteza badawcza: próba nie

jest losowa – ma wyraźną tendencję nielosową. Hipoteza zerowa: próba jest losowa.

Mediana wynosi Me = 150.4912 .

Sprawdzamy, które wyniki w próbie są niższe, a które wyższe od mediany. Mamy zatem, po kolei jak leci (symbol a: x < Me, symbol b: x > Me):

a, a, b, a, a, b, a, b, a, a, b, a, a, b, a, b, a, b, b, a, a, a, b, a, b, b, a, b, b, b, b, b, a, a, a,
b, b, b, a, a, b, b, b, a, a, a, b, b, b, a, b, b, a, a, a, b, a, a, a, b, b, b, b, a, b, a, b, a, b, a.

Liczba serii S = 39. Liczba symboli a = 35. Liczba symboli b = 35. Wszystkich liczb N = 70.

Dla liczb a lub b powyżej 20 stosujemy aproksymację do rozkładu normalnego. Liczymy estymator średniej i wariancji:

$$m\hat{} := 1 + ( 2 \cdot a \cdot b / N ) = 1 + (2 \cdot 35 \cdot 35 / 70 ) = 36;$$
$$s\hat{}^2 := 2 \cdot a \cdot b [ 2 \cdot a \cdot b – N ] / [ N^2 (N – 1) ] = 35 \cdot 34 / 69 = 17.2464; \; s\hat{} = 4.1529 .$$

Normalizujemy statystykę S – liczby serii; asympotytcznie jej rozkład dąży do rozkładu normalnego N(0,1). Mówiąc kolokwialnie – my sprawdzimy, jak daleko naszemu S do normalności, tj., czy nasza wartość Z (znormalizowane S) znajduje się w obszarze krytycznym N(0,1) dla poziomu ufności 95%.

$$Z := ( S – m\hat{} ) / s\hat{} = (39 – 36) / 4.1529 = 0.7224 .$$
(Prawostronna) granica obszaru krytycznego N(0,1) dla $\alpha = 0.05$ wynosi $z^* = 1.96$.

Jako, że nasza wartość Z znajduje się w jądrze rozkładu, a nie w dalekim skrzydle, odciętym przez wartość krytyczną (tzn. mamy w naszym przypadku $–z^* < Z < z^*$), **bpdo** hipotezy zerowej o losowości rozkładu. Rozkład wyników można uznać za **losowy.**

Ad 3.

M już policzyliśmy w 1.
$s^2 = 21.3193$ .
Mo (wartość modalna = dominanta = najczęstsza) nie istnieje – zmienna nie jest całkowita!
Me (medianę = środek ciężkości rozkładu) policzyliśmy w 2.
Ku (kurtoza) = – 0.0510 (rozkład mezokurtyczny = bardzo podobny do normalnego).
Sk (skośność) = – 0.0573 (niska = rozkład symetryczny).

Ad 4.

Uwaga wstępna: przedziały ufności szacuje się dla parametrów (w populacji), nie dla statystyk (w próbie).

Metodę szacowania przedziału ufności dla parametrów $\mu$ i $\sigma$ (odchylenie standardowe = pierwiastek z wariancji) w populacji można znaleźć np. w polskiej wikipedii. Gdyby zadania 4 i 5 były odwrócone kolejnością, znalibyśmy już rozkład normalny odniesienia (wzorcowy) dla naszej próby – co zresztą (przyjęcie założenia o podobieństwie próby do rozkładu normalnego) jest niezbędnym założeniem dla legalności poniższych obliczeń. Tu obliczenia wykonujemy póki co "na słowo honoru".

Możemy założyć, że nasza próba jest duża (umownie, N>30). A zatem, dla M:

$$pu = t(df = N-1) \cdot s / \sqrt{(N - 1)} = (około)\ 1.99 \cdot 4.6173 / 8.3066 = 1.1062,$$

gdzie: t = wartość rozkładu Studenta dla N – 1 stopni swobody;
s = odchylenie standardowe w próbie;
N = liczebność próby
pu = półprzedział ufności; ostatecznie przedział **[μ – pu; μ + pu]** szacuje przedział ufności.

Tymczasem dla σ:

$$pu = s / ( 1 + u(\alpha)/ \sqrt{2N} ) = 4.6173 / (1 + 1.96/11.8322) = 3.9611,$$

gdzie u(α) to wartość krytyczna z rozkładu N(0,1) dla danego α (= 0.05).
Przedział **[σ – pu; σ + pu] szacuje** przedział ufności dla odchylenia standardowego.

Ad 5.

Wykonujemy test Kołmogorowa-Smirnowa (K-S). Odpowie on na pytanie, czy możemy utrzymać założenie o podobieństwie naszego rozkładu do rozkładu normalnego o znajdowanych w wyniku testu parametrach. Istotność wyniku oznaczałaby konieczność odrzucenia tezy o podobieństwie; przeciwnie – nieistotność (bddo hipotezy zerowej) utrzymuje tezę o podobieństwie w mocy. Jest to rzadki przykład testu, w którym uczony liczy na nieobalenie hipotezy zerowej, a zamiast niej, badawczej – o wyraźnej różnicy między próbą, a rozkładem normalnym. Jest tak dlatego, iż hipotezy badawcze zawsze mówią o różnicach, odmienności, charakterystycznej tendencji – a hipotezy zerowe przeciwnie, o ich braku. Tu właśnie interesuje nas ten brak!

Kalkulacja jest o wiele zbyt skomplikowana, aby dokonywać ją ręcznie. Obliczamy za pomocą pakietu statystycznego SPSS 20:

**Analiza -> Testy nieparametryczne -> Testy tradycyjne -> K-S dla jednej próby...**

Z raportu **Test Kołmogorowa-Smirnowa dla jednej próby** odczytujemy:
Rozkład w próbie odpowiada rozkładowi normalnemu o parametrach: N(150.059, 4.617) .

Krytyczny poziom istotności asymptotycznej (dwustronnej) wychodzi 0.769, co znacznie przekracza istotność 0.05 (czyli: **wynik nieistotny**); wnioskujemy bpdo hipotezy zerowej o normalności rozkładu – tak jak chcieliśmy. Otrzymany wynik różni się jednak od N(150, 5). Można więc powiedzieć, że rozkład danych jest podobny do N(150.059, 4.617), a zatem **nie** jest podobny do N(150, 5).

Jak odpowiedzieć na pytanie: czy te dwa rozkłady są do siebie podobne, i na ile? Nie mam pojęcia i chyba nikt nie ma. Bowiem co miałoby być tego miarą? Bardziej jako ciekawostkę, niż próbę odpowiedzi, przytoczę fakt, iż zmienna, będąca różnicą dwóch zmiennych o rozkładach normalnych ma rozkład normalny; jego średnia jest równa różnicy średnich (w naszym przypadku: 5.000 – 4.617 = 0.383), oraz wariancja jest sumą wariancji obu rozkładów (czyli $\sigma^2 = 5.000^2 + 4.617^2 = 46.317 \Rightarrow s = 6.806$).

Ad 6.

Regresja liniowa oznacza odnalezienie równania funkcji liniowej ($y = a \cdot x + b$) , która najlepiej (najściślej) dopasowuje się do wyników doświadczalnych. Szukamy takiego współczynnika kierunkowego oraz wolnego wyrazu prostej, aby błąd dopasowania, definiowany jako suma kwadratów różnic pomiędzy wartościami funkcji liniowej a wartościami z próby (= suma reszt regresji), był najmniejszy.

Dodam, iż można także rozważać i rozwiązać regresję liniową wielu zmiennych

niezależnych, tj. taką, w której wartość zmiennej zależnej (y) zależy od *n* różnych zmiennych niezależnych $x_i$ w postaci:

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_n \cdot x_n + b .$$

My mamy tylko jedną zmienną niezależną: x = numer porządkowy.

Dla tej liczby punktów (N=70) wykonanie procedury ręcznie byłoby niezwykle mozolne (choć jest wykonalne). Chodzi o znalezienie ekstremów lokalnych w sensie minimum funkcji dwóch zmiennych a i b, jaką jest suma reszt regresji. W niektórych zastosowaniach liczy się minimum np. Funkcji wykładniczej od sumy reszt regresji – metodą konstrukcji najmniej obciążonych estymatorów a^ i b^. My sięgniemy po gotowy wynik z pakietu SPSS 20.

**Analiza -> Regresja -> Liniowa...**

Z raportu **Współczynniki** wynika $y = -0.021 \cdot x + 150.800$ .