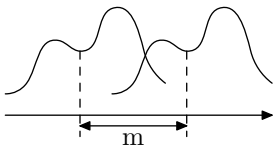
Коротко и строго о непараметрических тестах. (ver 26.02.08).
Текст можно скачать на www.xion.ru (учеба-2 курс-теория вероятностей).
Вопросы/комментарии/предложения можно смело отправлять на roah@yandex.ru (Борису Демешеву)

Sign test. Тест знаков.
Предпосылки:
X_i - независимы и имеют общую медиану m
Проверяемая гипотеза:
H_0: $m = m_0$, "медиана равна m_0"
H_a: $m > m_0$ или $m \neq m_0$ или $m < m_0$
Формула расчета:
B - количество наблюдений больше предполагаемой медианы m_0, число «плюсов».
Распределение при верной H_0:
Биномиальное (асимптотически нормальное)
$E(B) = \frac{n}{2}$, $Var(B) = \frac{n}{4}$

Пример.
Имеются наблюдения за говорливостью 30 попугаев (слов/день):
34, 56, 32, 45, 34, 45, 67, 1, 34, 12, 123, ... , 37 (всего 13 наблюдений меньше 40)
Проверить гипотезу о том, что медиана равна 40 (слов/день).
Решение.
Будем проверять H_0: $m = 40$ против H_a: $m \neq 40$ при $\alpha = 0.05$
$B = 17$. Поскольку число наблюдений велико, будем использовать нормальную аппроксимацию. $Z = \frac{B - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$, где $n = 30$ - число наблюдений.
Получаем, что $Z = 0,73$ и $Z_{critical} = 1.96$. Вывод: H_0 не отвергается, т.е. имеющиеся данные не противоречат гипотезе о том, что половина попугаев говорит более 40 слов в день, а половина - меньше.

Wilcoxon Rank Sum Test. Он же **Mann-Whitney test**.
Предпосылки:
X_i, $i = 1..n_1$ - независимы, имеют одинаковую функцию плотности $p_X(t)$.
Y_j, $j = 1..n_2$ - независимы, имеют одинаковую функцию плотности $p_Y(t)$.
Функции $p_X(t)$ и $p_Y(t)$ имеют одинаковую форму и отличаются только сдвигом влево-вправо, т.е. $p_Y(t + m) = p_X(t)$
Проверяемая гипотеза:
H_0: $m = 0$, функции плотности полностью совпадают, X и Y взяты из одной генеральной совокупности.
H_a: $m > 0$ или $m \neq 0$ или $m < 0$

Иллюстрация:



Формула расчета:
Наблюдения за X_i и Y_i сваливаем в одну кучу и группируем по возрастанию. Расставляем ранги (порядковые номера по возрастанию).
W_1 - сумма рангов всех X_i.
$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$

Распределение при верной H_0:
Особое, есть специальные таблицы. (асимптотически нормальное)
$E(W_1) = \frac{n_1(n_1+n_2+1)}{2}$, $Var(W_1) = \frac{n_1 n_2(n_1+n_2+1)}{12}$

Формула расчета (альтернативная):
Переберем все возможные пары (X_i, Y_j). И посчитаем, сколько раз оказалось, что $X_i > Y_j$. Обозначим это число U_1.
Можно например, в прямоугольной таблице (где по столбцам X_i, а по строкам Y_j) расставить «плюсики» и посчитать их количество.
Два способа расчета связаны соотношением: $U_1 = W_1 - \frac{n_1(n_1+1)}{2}$
Распределение при верной H_0:
Особое, есть специальные таблицы. (асимптотически нормальное)
$E(U_1) = \frac{n_1 n_2}{2}$, $Var(U_1) = \frac{n_1 n_2(n_1+n_2+1)}{12}$

Доказательство формул мат. ожидания и дисперсии
Следует отметить, что если H_0 верна, то W_1 - это сумма наугад выбранных n_1 чисел из набора 1, 2, 3, ..., $(n_1 + n_2)$
Общее количество наблюдений $N = n_1 + n_2$
Пусть W_1 - это сумма наугад выбранных n_1 чисел из набора 1, 2, 3, ..., N, а X_i - это i-ое выбранное число, т.е. $W_1 = X_1 + ... + X_{n_1}$.
$E(W_1) = n_1 \cdot E(X_1) = n_1 \cdot \left(1 \cdot \frac{1}{N} + ... + N \cdot \frac{1}{N}\right) = n_1 \cdot \frac{1+N}{2}$
И $E(U_1) = E(W_1) - \frac{n_1(n_1+1)}{2} = \frac{n_1 n_2}{2}$

$Var(U_1) = Var(W_1)$.
Воспользовавшись известной формулой подправки на конечность совокупности (см. Лемму 1 из Приложения), получаем:
$Var(W_1) = n_1 \cdot Var(X_1) \cdot \frac{N-n_1}{N-1}$, где $N = n_1 + n_2$.
$Var(X_1) = E(X_1^2) - (E(X_1))^2$.
$E(X_1^2) = \frac{1}{N} \cdot (1^2 + 2^2 + ... + N^2)$
По формуле для суммы квадратов (см. Лемму 2 из Приложения), получаем:
$E(X_1^2) = \frac{(N+1)(2N+1)}{6}$
$Var(X_1) = \frac{(N+1)(2N+1)}{6} - (\frac{1+N}{2})^2 = \frac{(N+1)(N-1)}{12}$
И в результате получаем, что:
$Var(W_1) = \frac{n_1 n_2(n_1+n_2+1)}{12}$

Пример:
Вашему вниманию представлены результаты прыжков в длину Васи Сидорова. Среди болельщиц присутствовала Аня Иванова (его первая любовь): 1,83; 1,64; 2,27; 1,78; 1,89; 2,33; 1,61; 2,31. Аня Иванова среди болельщиц не присутствовала: 1,26; 1,41; 2,05; 1,07; 1,59; 1,96; 1,29; 1,52; 1,18; 1,47.
С помощью теста (Mann-Whitney) проверьте гипотезу о том, что присутствие Ани Ивановой положительно влияет на результаты Васи Сидорова. Уровень значимости $\alpha = 0.05$.
Решение:
$n_1 = 8$, $n_2 = 10$.
Формулировка гипотез:
H_0: $m = 0$, Анино присутствие не сказывается на Васиных успехах
H_a: $m > 0$, Аня оказывает положительное воздействие
После упорядочивания получаем, что результаты прыжков, при которых Аня присутствовала, занимают 9, 10, 11, 12, 13, 16, 17 и 18 места. Соответственно, $W_1 = 106$ и $U_1 = 70$.
При верной H_0, получаем, что $E(U_1) = 40$ и $Var(U_1) = 126\frac{2}{3}$.
Для простоты будем пользоваться нормальной аппроксимацией.
Наблюдаемое $Z = 2.7$ при $Z_{critical} = 1.65$ (т.к. односторонняя область).
Вывод: H_0 отвергается в пользу H_a, т.е. присутствие Ани положительно сказывается на результатах

Васиных прыжков в длину.

Wilcoxon Signed Rank Test.
Предпосылки:
Имеются парные наблюдения X_i и Y_i, где $i = 1..n$.
Обозначим $d_i = (X_i - Y_i)$. d_i - независимы, функции плотности d_i симметричны относительно общего m. Сами функции плотности могут не совпадать. Среднее у разных X_i может быть разным, сами X_i могут быть зависимыми.
Проверяемая гипотеза:
H_0: $m = 0$, переход от X_i к Y_i не меняет среднего
H_a: $m > 0$ или $m \neq 0$ или $m < 0$

Примечание:
Возможен иной вариант применения этого теста:
Вариант предпосылок:
X_i - независимы и симметрично распределены относительно общего среднего m. В остальном законы распределения могут отличаться. В этом случае, $d_i = X_i - m_0$
Вариант проверяемой гипотезы:
H_0: $m = m_0$, "медиана (или среднее) равна m_0"
H_a: $m > m_0$ или $m \neq m_0$ или $m < m_0$
Формула расчета:
Упорядочиваем d_i по возрастанию абсолютной величины. Расставляем ранги (порядковые номера по возрастанию).
T^+ - сумма рангов положительных d_i.
Распределение при верной H_0:
Особое, есть специальные таблицы. (асимптотически нормальное)
$E(T^+) = \frac{n(n+1)}{4}$ и $Var(T^+) = \frac{n(n+1)(2n+1)}{24}$

Доказательство формул мат. ожидания и дисперсии:
Можно представить T^+ в виде:
$T^+ = 1 \cdot I_1 + 2 \cdot I_2 + ... + n \cdot I_n$, где I_k равно 1, если ранг k достался положительному d_i. Если H_0 верна, то $P(I_k = 1) = 0.5$.
Следовательно:
$E(T^+) = \frac{1}{2} \cdot (1 + 2 + ... + n) = \frac{n(n+1)}{4}$
Для дисперсии снова воспользуемся формулой для суммы квадратов (лемма 2 из Приложения):
$Var(T^+) = \frac{1}{4} \cdot (1^2 + 2^2 + ... + n^2) = \frac{n(n+1)(2n+1)}{24}$

Пример:
Некоторые результаты 2-х контрольных по теории вероятностей выглядят следующим образом (указан результат за вторую контрольную и в скобках результат за первую):
43(55), 113(108), 97(53), 68(42), 94(67), 90.5(97), 35(91), 126(127), 102(78), 89(83).
Можно ли считать (при $\alpha = 0.05$), что вторую контрольную написали лучше?
Решение:
Предположим, что изменения результатов имеют симметричные распределения относительно общего числа m.
Проверяем:
H_0: $m = 0$, в среднем результат такой же
H_a: $m > 0$, вторую контрольную написали лучше
Разницы d_i равны (упорядочены по модулю): -1, 5, 6, -6.5, -12, 24, 26, 27, 44, -56
Положительные d_i занимают места: 2, 3, 6, 7, 8, 9, значит $T^+ = 35$.
Если H_0 верна, то $E(T^+) = 27.5$ и $Var(T^+) = 96.25$ ($n = 10$).
Для простоты используем нормальное распределение:
Получаем наблюдаемое $Z = 0.76$ при $Z_{critical} = 1.65$

Вывод: имеющиеся наблюдения не противоречат H_0. Т.е. первую и вторую контрольную написали в целом одинаково.

Run's test. Тест серий.
Предпосылки:
Имеются результаты N последовательных испытаний. Каждое из них это успех (+) или неуспех (−). Имеется n_+ успехов и n_- неуспехов.
Проверяемая гипотеза:
H_0: Испытания независимы.
H_a: Испытания зависимы, причем зависимость сказывается на ожидаемом числе серий. (Строгая ли???)
Формула расчета:
Считаем T число серий (серия - это последовательность из одинаковых знаков). Число серий равно числу смен знака плюс единица.
Распределение при верной H_0:
Особое, есть специальные таблицы. (асимптотически нормальное)
$E(T) = \frac{2n_+n_-}{n_++n_-} + 1$ и $Var(T) = \frac{2n_+n_-(2n_+n_--N)}{N^2(N+1)}$

Доказательство формул мат. ожидания и дисперсии:
Обозначим $N = n_+ + n_-$. Если верна H_0, то можно представить себе следующий эксперимент: в корзине перемешиваются n_+ успехов и n_- неуспехов, затем их достают из корзины по очереди.
Разложим T в сумму:
$T = I_1 + I_2 + ... + I_N$, где I_k равно 1, если на k-ом числе происходит начало новой серии (смена знака). При этом $I_1 = 1$ тождественно.
Далее пригодятся несколько вспомогательных результатов:
$E(I_2) = \frac{n_+}{N} \cdot \frac{n_-}{N-1} + \frac{n_-}{N} \cdot \frac{n_+}{N-1} = \frac{2n_+n_-}{N(N-1)}$ и $E(I_2^2) = \frac{2n_+n_-}{N(N-1)}$.
Для определения $E(I_iI_j)$ важно знать расстояние между i и j.
$E(I_2I_3) = \frac{n_+}{N}\frac{n_-}{N-1}\frac{n_+-1}{N-2} + \frac{n_-}{N}\frac{n_+}{N-1}\frac{n_--1}{N-2} = \frac{n_+n_-}{N(N-1)}$.
$E(I_2I_4) = P(I_2 = 1) \cdot P(I_4 = 1|I_2 = 1) = \frac{2n_+n_-}{N(N-1)}\frac{2(n_+-1)(n_--1)}{(N-2)(N-3)}$
Теперь легко находим:
$E(T) = 1 + (N-1) \cdot E(I_2)$, и, следовательно, $E(T) = 1 + 2\frac{n_+n_-}{N}$.
$Var(T) = Var(I_2 + I_3 + ... + I_N) =$
$= (N-1)Var(I_2) + 2(N-2)Cov(I_2, I_3) + (N-2)(N-3)Cov(I_2, I_4) =$
$= (N-1)E(I_2^2) + 2(N-2)E(I_2I_3) + (N-2)(N-3)E(I_2I_4) - (N-1)^2(E(I_2))^2 =$
После подстановки и упражнения по алгебре 9-го класса, получаем:
$Var(T) = \frac{2n_+n_-(2n_+n_--N)}{N^2(N+1)}$

Пример:
Садовник осматривал по очереди розовые кусты вдоль ограды. Всего вдоль ограды растет 30 розовых кустов. Из них оказалось 20 здоровых и 10 больных.
Вот заметки садовника: $+++\ominus++\ominus\ominus\ominus++++\ominus\ominus++\ominus++++++\ominus\ominus\ominus+++$
(+ - здоровый куст, $\ominus$ - больной куст)
а) С помощью теста серий проверьте гипотезу о независимости испытаний
б) Какой естественный смысл имеет эта гипотеза?
Подсказка: можно использовать нормальное распределение
Решение:
Запишем условие в наших обозначениях: $T = 11$, $n_+ = 20$, $n_- = 10$, $N = 30$
Следовательно, $E(T) = 14.33$, $Var(T) = 5.30$
Выбираем уровень значимости 5%
Какой должна быть альтернативная гипотеза?
В данном случае логично предположить, что альтернативной гипотезой является «заразность» заболевания, т.е. имеем одностороннюю область, где H_0 отвергается

Используем нормальное распределение. $Z = -1.45$, $Z_{critical} = -1.64$
Вывод H_0 - не отвергается.

Приложение.

Лемма 1. Подправка на конечность совокупности.
Пусть имеется N чисел, из которых наугад выбираются n.
Обозначим X_i - i-ое извлекаемое число, S_n - сумму извлеченных чисел, $\overline{X}_n$ - среднее арифметическое извлеченных чисел и $\sigma^2 = Var(X_i)$.
В таких обозначениях:
$Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$ при $i \neq j$.
$Var(\overline{X}_n) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$.
$Var(S_n) = n \cdot \sigma^2 \cdot \frac{N-n}{N-1}$.
Доказательство:
Заметим, что $\sum_{i=1}^{N} X_i = const$, поэтому $Cov(X_1, X_1 + X_2 + ... + X_N) = 0$.
Воспользуемся тем, что $Cov(X_1, X_2) = Cov(X_1, X_3) = ... = Cov(X_1, X_N)$.
Таким образом мы получаем $Var(X_1) + (N-1)Cov(X_1, X_2) = 0$.
И, следовательно, $Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$ для $i \neq j$.
Значит $Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\left(n \cdot \sigma^2 + 2 \cdot C_n^2 \cdot \left(-\frac{\sigma^2}{N-1}\right)\right) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
Осталось заметить, что $Var(S_n) = n \cdot \sigma^2 \cdot \frac{N-n}{N-1}$

Лемма 2.

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Доказательство:
Для $n = 1$ формула верна.
Пусть она верна для некоторого n. Докажем ее для $(n+1)$.

$$\sum_{i=1}^{n+1} i^2 = \frac{n(n+1)(2n+1)}{6} + (n+1)^2 = ... = \frac{[n+1]([n+1]+1)(2[n+1]+1)}{6}$$

P.S.
Современная педагогика зародилась в Абхазии, где высоко в горах одаренные чабаны могли часами удерживать аудиторию в пять, а то и в шесть тысяч баранов.