Коротко о некоторых параметрических тестах. (ver 01.08.06).
Текст можно скачать на www.xion.ru (учеба-2 курс-теория вероятностей).
Вопросы/комментарии/предложения/найденные ошибки можно смело отправлять на roah@yandex.ru (Борису Демешеву)

Quote
Во избежание несчастных случаев торпеды хранить так, чтобы верхняя их часть находилась внизу, а нижняя наверху. Дабы персонал не путал верх с низом, на верхней части торпеды сделать надпись «верх»

из инструкции

Обозначения

$\bar{X} = \frac{X_1+...+X_n}{n}$ - несмещенная оценка математического ожидания
$\hat{\sigma}^2 = \frac{\sum_i (X_i-\bar{X})^2}{n-1}$ - несмещенная оценка дисперсии
$\hat{v}^2 = \frac{\sum_i (X_i-\bar{X})^2}{n}$ - оценка дисперсии, при больших n слабо отличается от $\hat{\sigma}^2$

Случай 0-1
Часто рассматривается случай, когда X_i принимают только значения 0 и 1. Значение 1 с вероятностью p, значение 0 с вероятностью $q = 1 - p$.
В этом случае:
1. $E(X_i) = p$, $Var(X_i) = pq = p(1-p)$.
2. Вместо $\bar{X}$ часто используется обозначение $\hat{p}$
3. $\hat{v}^2 = \hat{p}(1-\hat{p})$
Доказательство 3:
$\sum (X_i - \bar{X})^2 = \sum (X_i^2 + \bar{X}^2 - 2X_i\bar{X}) =$
$= \sum X_i + n\bar{X}^2 - 2\bar{X}\sum X_i = n\bar{X} + n\bar{X}^2 - 2n\bar{X}^2 = n\bar{X} - n\bar{X}^2 = n\hat{p}(1-\hat{p})$

Definitions
Определение 1.
Распределение случайной величины K называется χ^2 распределением с n степенями свободы, если величину можно представить в виде $K = Z_1^2 + ... + Z_n^2$, где Z_i iid, $N(0;1)$
Т.е. есть χ^2-распределение номер 1, есть χ^2-распределение номер 2 и т.д.
Определение 2.
Распределение случайной величины T называется t-распределением с n степенями свободы, если величину можно представить в виде $T = \frac{Z}{\sqrt{\frac{K}{n}}}$, где $Z \sim N(0;1)$, $K \sim \chi^2_n$, Z и K независимы

Определение 3.
Распределение случайной величины F называется F-распределением с n, k степенями свободы, если величину можно представить в виде $F = \frac{X/n}{Y/k}$, где $X \sim \chi^2_n$ и $Y \sim \chi^2_k$ и X и Y независимы

Одна выборка

Асимптотический результат.
Если:
1. X_i - независимы, одинаково распределены,

2. $E(X_i) = \mu$, $Var(X_i) = \sigma^2$,
То:
1. $Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$
2. $Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$
Примечание:
В пункте 2 вместо $\hat{\sigma}^2$ можно взять любую другую состоятельную оценку дисперсии, например $\hat{v}^2$

Точный результат.
Добавив дополнительное условие нормальности отдельных X_i получаем:
Если:
1. X_i - независимы, одинаково распределены,
2. $E(X_i) = \mu$, $Var(X_i) = \sigma^2$,
3. X_i - нормально распределены,
То:
1. $Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$ имеет нормальное распределение $N(0;1)$
2. $Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$ имеет t-распределение с $n-1$ степенью свободы
3. $Q = \frac{(n-\hat{1})\sigma^2}{\sigma^2}$ имеет χ^2 распределение с $n-1$ степенью свободы.

Частный случай асимптотического результата.
Отдельно рассмотрим случай 0-1:
Если:
1. X_i - независимы, одинаково распределены,
2. X_i принимает значение 1 с вероятностью p и значение 0 с вероятностью $(1-p)$
То:
0. $E(X_i) = p$, $Var(X_i) = p(1-p)$, $\hat{v}^2 = \hat{p}(1-\hat{p})$, вместо $\bar{X}$ используем $\hat{p}$
1. $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$
2. $Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$

Две выборки

Если:
1. X_i - одинаково распределены между собой,
2. $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$,
3. Y_i - одинаково распределены между собой,
4. $E(Y_i) = \mu_y$, $Var(Y_i) = \sigma_y^2$,
5. Все случайные величины X_i и Y_j независимы.
То:
1. $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$
2. $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$
3. $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{(n_x-1)\hat{\sigma}_x^2 + (n_y-1)\hat{\sigma}_y^2}{n_x+n_y-2} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$ имеет асимптотически (т.е. при $n \to \infty$) нормальное распределение $N(0;1)$

Если:
1. X_i - одинаково распределены между собой,
2. $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$,
3. Y_i - одинаково распределены между собой,
4. $E(Y_i) = \mu_y$, $Var(Y_i) = \sigma_y^2$,
5. Все случайные величины X_i и Y_j независимы.
6. Все случайные величины X_i и Y_j нормально распределены

То:
1. $t = \frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{(n_x-1)\hat{\sigma}_x^2+(n_y-1)\hat{\sigma}_y^2}{n_x+n_y-2}\cdot\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$ имеет t-распределение с $(n_x + n_y - 2)$ степенью свободы
2. $F = \frac{\hat{\sigma}_x^2/\sigma_x^2}{\hat{\sigma}_y^2/\sigma_y^2}$ имеет F распределение с $(n_x - 1)$ и $(n_y - 1)$ степенями свободы.

Summary

Для проверки гипотез о среднем:
Если точный закон распределения X_i неизвестен или доподлинно известно, что он отличается от нормального (случай 0-1), то используется асимптотически нормальное распределение. Т.е. требуется большое n.
Если известно, что X_i нормальны, то (при любых n):
а) если известна оценка дисперсии $\hat{\sigma}^2$ - используем t. При больших n t-распределение перестанет отличаться от нормального.
б) если известна точная дисперсия (что бывает редко) - используем нормальное распределение
в) можно проверять гипотезы о дисперсии с помощью χ^2

Про χ^2 распределение

Проверка гипотезы о соответствии наблюдаемых частот заданному закону распределения

Предполагаемые частоты: p_i
Выборочные частоты: $\hat{p}_i$

Статистика (имеет χ^2 распределение с $(c-1)$ степенями свободы): $K = \sum \frac{(n\hat{p}_i-np_i)^2}{np_i}$
Можно, конечно, считать по иному:
$K = n\sum \frac{(\hat{p}_i-p_i)^2}{p_i}$
Или, если напрячь немного арифметический мускул:
$K = n\left(\sum \frac{\hat{p}_i^2}{p_i} - 1\right)$

Проверка гипотезы о независимости двух признаков

Выборочные частоты: $\hat{p}_{ij}$
Частоты, рассчитанные в предположении независимости: p_{ij}

Способ расчета эталонных «независимых» частот:
Сначала считаем оценки вероятностей по каждому признаку:
$p_{i\cdot} = \sum_j \hat{p}_{ij}$
$p_{\cdot j} = \sum_i \hat{p}_{ij}$
Если события A и B независимы, то $P(A \cap B) = P(A)P(B)$:

$p_{ij} = p_{\cdot j} p_{i\cdot}$

Статистика (имеет χ^2 распределение с $(r-1)(c-1)$ степенями свободы): $K = \sum \frac{(n\hat{p}_{ij} - np_{ij})^2}{np_{ij}}$
Или: $K = n \sum \frac{(\hat{p}_{ij} - p_{ij})^2}{p_{ij}}$
Или: $K = n \left(\sum \frac{\hat{p}_{ij}^2}{p_{ij}} - 1 \right)$

- Сколько степеней свободы?
- Столько, сколько вероятностей можно расставить «от фонаря» при соблюдении ограничений
В первом случае:
В обеих табличках (для p_i и $\hat{p}_i$) имеется c ячеек.
Действует единственное общее ограничение $\sum p_i = 1$ (и $\sum \hat{p}_i = 1$)
Значит, $(c-1)$ вероятность может быть любой, а последняя считается исходя из того, что сумма вероятностей равна 1.

Во втором случае:
В обеих табличках (для p_{ij} и $\hat{p}_{ij}$) имеется $r \cdot c$ ячеек
Общие ограничения: сумма вероятностей по каждой строке и по каждому столбцу должна быть одинакова для p_{ij} и для $\hat{p}_{ij}$.
Значит в каждом столбце кроме последнего можно поставить «от фонаря» $r-1$ число. Последний столбец рассчитается сам собой. Следовательно, получается $(r-1)(c-1)$ степень свободы.

Большинство авторов использует в формуле не частоты, а количества. Мне этот подход кажется менее удачным, потому, что нужно объяснять, что такое «эталонное» количество. При этом более туманным (imho) становится вычисление эталонного количества для проверки гипотезы о независимости признаков.
Переход от вероятностей к количествам прозрачен:
Эталонное количество: $E_i = np_i$
Выборочное количество: $V_i = n\hat{p}_i$
Соответственно меняются формулы.

Some proofs
Теорема 1.
Пусть X_i iid, $N(\mu; 1)$. Величину $\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ можно представить в виде $\sum_{i=1}^{n-1} Z_i^2$, где $Var(Z_i) = 1$ и $Cov(Z_i, Z_j) = 0$ для $i \neq j$.

Доказательство:
В качестве Z_k возьмем $\frac{\sum_{i=1}^{k} X_i - kX_{k+1}}{\sqrt{n(n+1)}}$
То, что $Var(Z_i) = 1$ и $Cov(Z_i, Z_j) = 0$ проверяется "в лоб".
Остается убедится в том, что $\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \sum_{i=1}^{n-1} Z_i^2$
Доказательство (предложила Алина Дурдыева)

Для $n = 1$ формула верна.
Пусть для некоторого n формула верна, т.е. $\sum_{i}^{n} (X_i - \bar{X}_n)^2 = \sum_{i}^{n-1} Z_i^2$

$\sum_{i=1}^{n+1} (X_i - \bar{X}_{n+1})^2 = \sum_{i=1}^{n+1} (X_i - \frac{X_{n+1} + n\bar{X}_n}{n+1})^2 = \sum_{i=1}^{n+1} (X_i - \bar{X}_n - \frac{X_{n+1} - \bar{X}_n}{n+1})^2 =$

$= \sum_{i=1}^{n}(X_i - \bar{X}_n - \frac{X_{n+1}-\bar{X}_n}{n+1})^2 + (X_{n+1} - \bar{X}_n - \frac{X_{n+1}-\bar{X}_n}{n+1})^2$

Обозначим $b = \frac{X_{n+1}-\bar{X}_n}{n+1}$.

$\sum_{i=1}^{n+1}(X_i - \bar{X}_{n+1})^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n - b)^2 + ((n+1)b - b))^2 =$
$= \left[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + nb^2 - 2b\sum_{i=1}^{n}(X_i - \bar{X}_n)\right] + n^2b^2 =$
$= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + b^2n(n+1)$
Осталось вспомнить, что $Z_n = \frac{\sum_{i=1}^{n} X_i - nX_{n+1}}{\sqrt{n(n+1)}}$.
Получаем, что $\sum_{i=1}^{n+1}(X_i - \bar{X}_{n+1})^2 = \sum_{i=1}^{n-1} Z_i^2 + Z_n^2$.

Следствие.
Пусть X_i iid, $N(\mu; \sigma^2)$. Из теоремы и определения нетрудно видеть, что:
$\frac{\sum_i^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2_{n-1}$
Если $\hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$, то $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$

Теорема 2.
Если X_i iid, $N(\mu_x; \sigma^2)$ и Y_i iid, $N(\mu_y; \sigma^2)$.
То: $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{(n_x-1)\hat{\sigma}_x^2 + (n_y-1)\hat{\sigma}_y^2}{n_x+n_y-2} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim t_{n_x+n_y-2}$
Доказательство:
Вспомним, что $\frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n_x-1}$ и $\frac{\sum(Y_j - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n_y-1}$.
Следовательно, $\frac{\sum(X_i - \bar{X})^2 + \sum(Y_j - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$.
Также известно, что $Var(\bar{X} - \bar{Y}) = \sigma^2 \cdot (\frac{1}{n_x} + \frac{1}{n_y})$.
Следовательно, $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma^2 \cdot (\frac{1}{n_x} + \frac{1}{n_y})}} \sim N(0; 1)$.
По определению, $t_k = \frac{N(0;1)}{\sqrt{\frac{\chi^k}{k}}}$
Получаем, что:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{(n_x-1)\hat{\sigma}_x^2 + (n_y-1)\hat{\sigma}_y^2}{n_x+n_y-2} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim t_{n_x+n_y-2}$$

Можно получить более простой асимптотический результат.
Заметим, что $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{Var(\bar{X} - \bar{Y})}} \sim N(0; 1)$.
$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$.
Заменим настоящие (неизвестные) дисперсии, на их оценки:
$\hat{\sigma}^2_{(\bar{X}-\bar{Y})} = \frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}$.
Получим асимптотический результат:
$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \to N(0; 1)$.

Про упрямство и эконометрику
Я упрямо обозначаю несмещенную оценку дисперсии знаком $\hat{\sigma}^2$, а не s^2, как большинство европейских авторов.
Почему?
В курсе эконометрики решается задача нахождения $\hat{\beta}_1$ и $\hat{\beta}_2$, если
$Y_i = \beta_1 + \beta_2 X_i + u_i$

Попутно находится оценка дисперсии u_i. Она равна $\hat{\sigma}^2 = \frac{RSS}{n-2}$.
Там оценивается два параметра, поэтому $(n-2)$. А здесь мы оцениваем один параметр - мат. ожидание Y_i:
$Y_i = \beta_1 + u_i$
МНК даст $\hat{\beta}_1 = \bar{Y}$, а оценка дисперсии будет $\hat{\sigma}^2 = \frac{RSS}{n-1}$