Basic Tests

Nguyen Quang Vinh - Nguyen Thi Tu Van

August 02, 2015

Outline

Chi-square (χ^2)

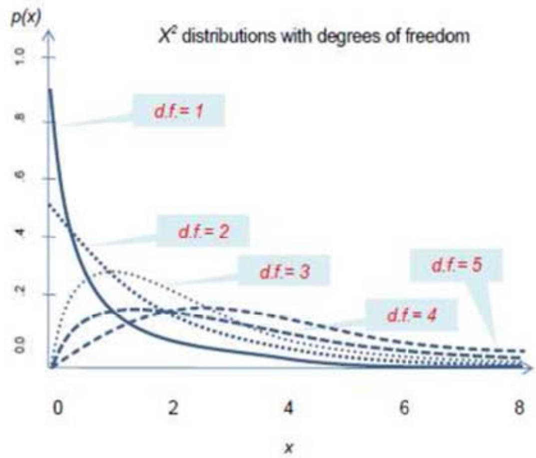
- One of the most widely used directly or indirectly distributions
- Testing hypothesis where data in form of frequencies: to test differences between proportions
- Most appropriate for use with categorical variables

Some Characteristics (opt.)

- The χ^2 distribution has one parameter, its $d.f.$ (k)
- It has a positive skew; the skew is less with more $d.f.$

- 1. Mean $= k$
  Variance $= 2k$
  Modal value $= k - 2$ (when $k \geq 2$) & $= 0$ (when $k = 1$)
  Median $\approx k - 0.7$
- 2. Shape: $k = 1$ & $k = 2$ vs. $k > 2$
- 3. Values range: $[0, +\infty)$
- 4. Sum of 2 or more independent χ^2 variables follows a χ^2 distribution

Chi-square distributions
with different degrees of freedom



Applications of the χ^2 Statistic

Observed frequencies (OBSERVATION) vs. Expected frequencies (HYPOTHESIS):

- (1) Test of goodness-of-fit
- (2) Test of independence
- (3) Test of homogeneity *(test of independence with fixed marginal totals)*

χ^2 test of goodness-of-fit

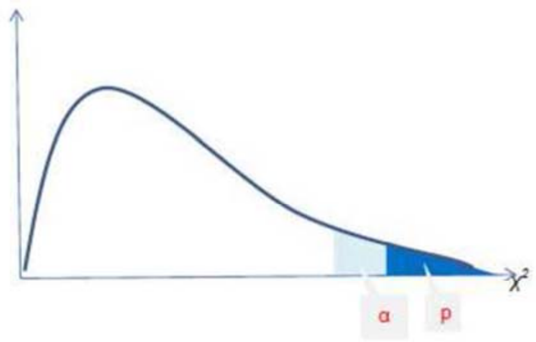
- A two-tailed test on p
- *For Binomial situation* :
  - $H_O : p = p_0$
  - $H_A : p \neq p_0$
- *For Multinomial situation*:(*)
  - $H_O : p_1 = p_{1_0},\ p_2 = p_{2_0},\ \ldots,\ p_k = p_{k_0}$
  - H_A : at least one of the $p_i's$ is incorrect

(*)Using χ^2 test of goodness-of-fit to test all of the proportions at once is better than using z tests to test proportions individually *(problem of the overall significance level)*.

χ^2 test of goodness-of-fit

Test statistic: $\chi^2_c = \sum \frac{(O-E)^2}{E}$
df= number of categories - 1
reject H_O if $\chi^2_c > \chi^2_{\alpha,df}$

Rejection area



χ^2 test of goodness-of-fit (opt.)

Testing H_O: $p = p_0$ vs. H_A: $p \neq p_0$

- Using z test:
  $z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
  reject H_O if $|z_c| > z_{1-\frac{\alpha}{2}}$
- Using χ^2 test:
  $\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$
  reject H_O if $\chi^2 > \chi^2_{\alpha,df=1}$

χ^2 test of goodness-of-fit

- How well the distribution of sample data conforms to some theorical distribution.(*)
- d.f. = k − r
- Small expected frequencies: there is disagreement among writers: 10, 5, 1 (*Cochran*).
  - Combining adjacent categories to achieve the suggested minimum.
  - When combining →↓number of categories→↓d.f.

(*)Kolmogorov-Smirnov test for continuous distribution.

χ^2 test of independence

- Most frequent use of χ^2 distribution
- A ***single*** population, where each member was classified according to 2 criteria:
  1^{st} *criteria* : row
  2^{nd} *criteria* : column
- Contingency table: r rows, c columns
- H_O : 2 criteria of classification are independent
  H_A : 2 criteria of classification are not independent
- df = (r − 1)(c − 1)

χ^2 test of independence

Small expected frequencies

- Small expected frequencies:
  df > 2 & no more than 20% of expected frequencies < 5 →1
  df < 30 → 2
  $n \geq 40$ → 1
- χ^2 test should not be used if:
  $n < 20$, or
  $20 \leq n < 40$ & any $E_i < 5$

χ^2 test of independence

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \ldots$$

reject H_O if $\chi^2_c > \chi^2_{\alpha, df=(r-1)(c-1)}$

χ^2 test of independence

2x2 table	Column 1	Column 2	Total
Row 1	a	b	a+b
Row 2	c	d	c+d
Total	a+c	b+d	n

$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$

$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$

Reject H_O if: $\chi^2 > \chi^2_{\alpha,1}$

$(df = (r-1)(c-1) = (2-1)(2-1) = 1)$

χ^2 test of independence (opt.)

$\chi^2_{corrected} = \sum \frac{(|O-E|-.5)^2}{E} = \frac{(|O_1-E_1|-.5)^2}{E_1} + \frac{(|O_2-E_2|-.5)^2}{E_2}$

$\chi^2_{corrected} = \frac{n(|ad-bc|-.5n)^2}{(a+b)(a+c)(b+d)(c+d)}$

Reject H_O if $\chi^2_{corrected} > \chi^2_{\alpha,1}$

Pro and Cons

χ^2 test of homogeneity

χ^2 test of independence with Fixed Marginal Totals

- To determine whether the distinct populations can be viewed as belonging to the same population (in terms of the criteria).

χ^2 test of homogeneity vs. χ^2 test of independence

- χ^2 test of independence: row and column totals are not under the control of the investigator
  χ^2 test of homogeneity: either row or column totals may be under the control of the investigator
- χ^2 test of independence: ? independent (the 2 criteria)
  χ^2 test of homogeneity: ? homogeneous (the samples drawn from the same population
- χ^2 test of homogeneity & χ^2 test of independence are mathematically equivalent but conceptually different.

χ^2 test of homogeneity (opt.)

- χ^2test of Homogeneity for the 2-sample case provides an alternative method for testing the H_O that: 2 population proportions are equal.
- A method for comparing of 2 population proportions using z statistic with pool proportion($\bar{p}$):
  Let $\hat{p}_1 = \frac{x_1}{n_1}$; $\hat{p}_2 = \frac{x_2}{n_2}$; $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$
  Test statistic: $z_c = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)_0}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1}+\frac{\bar{p}(1-\bar{p})}{n_2}}}$
- Note: $z_c = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$ is just for discussion purposes only. This equation should never be used as the test statistic for the difference between 2 proportions.

Fisher's exact test

- When expected value in χ^2test statistic is small.

	Treatment	Control	Total
O+	x	$K-x$	K
O-	$n-x$	$(N-K)-(n-x)$	N - K
Total	n	N-n	N

$$N \rightarrow \begin{Bmatrix} K & x \\ N-K & n\text{-}x \end{Bmatrix} \leftarrow n$$

$$P(x) = \frac{{}_K C_x \cdot {}_{N-K} C_{n-x}}{{}_N C_n}$$

Example

We have a result from a trial as follow:

	Treatment	Control	Total
O+	6	1	**7**
O-	2	4	**6**
Total	**8**	**5**	**13**

Listing all possible tables in the sample of size 13, which have:
7 positive outcomes & 8 subjects in treatment group
$\rightarrow$ We have 6 tables as follow:

	Treatment	Control	Total
O+	7	0	**7**
O-	1	5	**6**
Total	**8**	**5**	**13**

$$P(x=7) = \frac{{}_7C_7 \cdot {}_6C_1}{{}_{13}C_8} = \frac{6}{1287} = .0047$$

	Treatment	Control	Total
O+	6	1	**7**
O-	2	4	**6**
Total	**8**	**5**	**13**

$$P(x=6) = \frac{{}_7C_6 \cdot {}_6C_2}{{}_{13}C_8} = .0816$$

	Treatment	Control	Total
O+	5	2	**7**
O-	3	3	**6**
Total	**8**	**5**	**13**

$$P(x=5) = \frac{{}_7C_5 \cdot {}_6C_3}{{}_{13}C_8} = .3262$$

	Treatment	Control	Total
O+	4	3	**7**
O-	4	2	**6**
Total	**8**	**5**	**13**

$$P(x=4) = \frac{{}_7C_4 \cdot {}_6C_4}{{}_{13}C_8} = .4070$$

	Treatment	Control	Total
O+	*3*	4	**7**
O-	5	1	**6**
Total	**8**	**5**	**13**

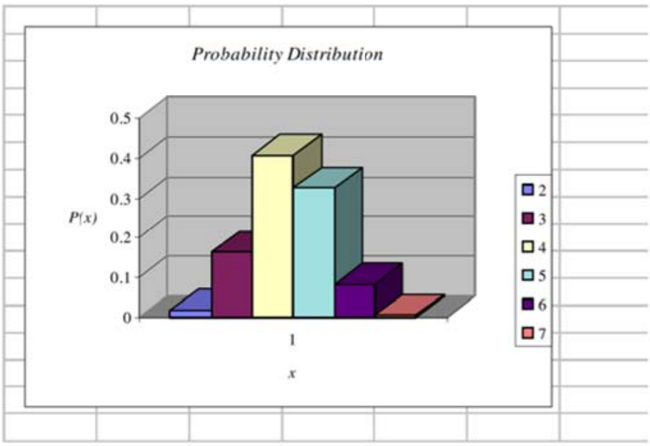
$$P(x=3) = \frac{{}_7C_3 \cdot {}_6C_5}{{}_{13}C_8} = .1632$$

	Treatment	Control	Total
O+	*2*	5	**7**
O-	6	0	**6**
Total	**8**	**5**	**13**

$$P(x=2) = \frac{{}_7C_2 \cdot {}_6C_6}{{}_{13}C_8} = .0163$$

*A useful check is that all the probabilities should sum to one (within the limits of rounding)

Probability distribution



Hypothesis

- $H_O : \pi_T = \pi_C$
  (no difference between treatment & control group)
- $H_A : \pi_T > \pi_C$ (1-tailed), or
- $H_A : \pi_T \neq \pi_C$ (2-tailed)

Calculate p value

- The observed set has a probability of 0.0816
- The p value is the probability of getting the observed set, or one more extreme.
- One tailed p value:
  - (1) p(x≥6)= p(x=6) + p(x=7)= 0.0816 + 0.0047= 0.0863 (this is the conventional approach).
  - (2) Armitage & Berry (1994) favor the mid p value: 0.5 x 0.0816 + 0.0047 = 0.0455
- Two tailed p value:
  - (1) p(x≥6 or x≤2)= p(x=2)+ p(x=6)+ p(x=7) = 0.0816 + 0.0047 + 0.0163 = 0.1026
  - (2) Double the one tailed result *(approximation)*, thus:
    p= 2 x 0.0863 = 0.1726 (for the conventional approach) or
    p= 2 x 0.0455 = 0.091 (for the mid P approach)

The conventional vs. the mid p (opt.)

- The conventional approach to calculating the p value for Fisher's exact test has been shown to be conservative (that is, it requires more evidence than is necessary to reject a false H_O)
- The mid P is less conservative (that is more powerful) & also has some theoretical advantages

Why is Fisher's test called an exact test? (opt.)

- Because of the discrete nature of the data, and the limited amount of it, combinations of results which give the same marginal totals can be listed, and probabilities attached to them.
  → thus, given these marginal totals we can work out exactly what is the probability of getting an observed result.

The t distribution

◇ PROBLEM:

- σ is known & not known μ (!)
- Indeed, it is the usual case, σ & μ is unknown

◇ We cannot make use the statistic: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ because σ is unknown, even when n is large,

→ use $s = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$ to replace σ

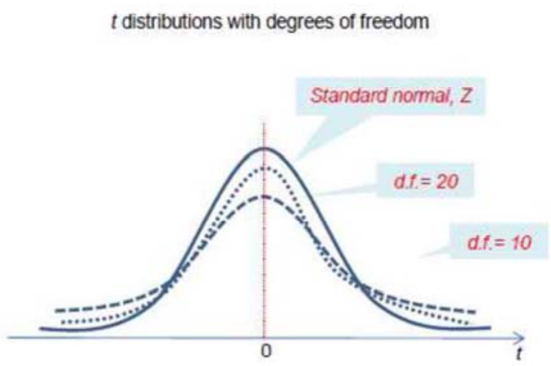
The t distribution

- William Sealy Gosset "Student" (1908) → Student's t distribution = t distribution.
- The quantity: $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ follows this distribution.

The t distribution (opt.)

- 1. It has a mean of 0.
- 2. It is symmetrical about the mean.
- 3. Variance: In general, it has a variance greater than 1, but the variance approaches 1 as the sample size becomes large. For $\nu > 2$, the variance of the t distribution is $\frac{\nu}{\nu-2}$ $\Longleftrightarrow$ For $n > 3$, the variance of the t distribution is $\frac{n-1}{n-3}$
- 4. The variable t ranges from $-\infty$ to $+\infty$
- 5. The t distribution = a family of distributions, since there is a different distribution for each sample value of $\nu = n - 1$
- 6. Compared to the normal distribution the t distribution is less peaked in the center & has higher tails
- 7. The t distribution approaches the normal distribution as n - 1 approaches infinity.

t distributions with degrees of freedom



Notice

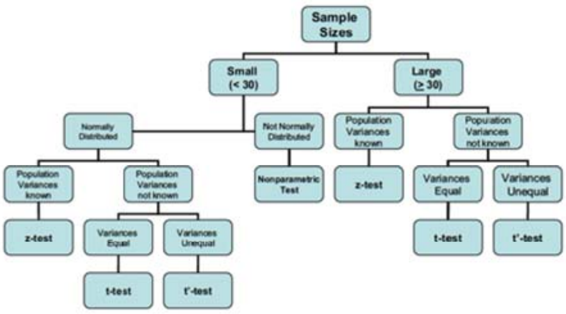
A requirement for valid use of the t distribution: sample must be drawn from
☼ a normal distribution. or
☼ at least, a mound-shaped distribution.

An interval estimate

- In general, an interval estimate is obtained by the formula: *estimator ± (reliability coefficient) x (standard error)*
- What is different is the source of the reliability coefficient:
  - In particular, when sampling is from a normal distribution with known variance, an interval estimate for μ may be expressed as: $\overline{x} \pm z_{\frac{\alpha}{2}} \sigma_{\overline{x}}$
  - when sampling is from a normal distribution with unknown variance, the $100(1-\alpha)\%$ confidence interval estimate for the population mean, μ, is given by: $\overline{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$

z, t or t'

Deciding between z, t, or t'



Flowchart for use in deciding whether the reliability factor should be use z, t, or t' when making inferences about the difference between two population means (* use a nonparametric procedure)

The Mann-Whitney Test

- When small samples from suspected nonnormal population - substitution of 2-sample t test.
- Assumptions for M-W test:
  1. Independent, Random samples
  2. Data at least ordinal
  3. If the 2 populations differ, they differ only in location (e.g., the 2 populations have the same variance and shape).
- Hypothesis: H_O : 2 populations have identical of the probability distribution *vs.*
  H_A : 2 populations differ in location (2-tailed), or
  H_A : population 1 is shifted to the right of population 2 (1-tailed), or
  H_A : population 2 is shifted to the right of population 1 (1-tailed)

Correlation & Regression

- Nature & strength of the relationship between 2 variables: BP & age, cholesterol & age, heigth & weight, size & weight of fetus, drug & heart rate
  → Correlation & Regression analysis
- Correlation: the strength of the association between 2 variables.
  Correlation refers to the interdependence or co-relationship of variables.
- Regression: predict, or estimate.
  Regression is a way of describing how one variable, the outcome, is numerically related to predictor variable(s).

Data types for correlation/regression analysis

- Need our data to be quantitative / continuous / numerical.
- Basic test: If data can meaningfully be portrayed on a scatter plot.

Pearson's correlation

- Pearson correlation detect linear relationships between variables.

Pearson correlation coefficient
(Pearson's product moment correlation coefficient)

- Pearson's correlation coefficient is a measure of the closeness linear association between X and Y.
- Denoted by r (sample statistic), and ρ (population parameter).
- Won't go into calculations for r (understand what it means).

Interpretation of r

r is a much abused statistic

- -1 < r < 1
- Sign of + or −.
- Value r doesn't mean the steepness of the slope.

Interpretation of r

r is a much abused statistic

- The large |r| is, the stronger is the linear relationship.
  + Values of r close to −1 or +1 indicate a strong (negative or positive) linear relationship.
  r is close to ± 1 then this does NOT mean that there is a good causal relationship between X and Y. It shows only that the sample data is close to a straight line.
  + Values of r close to zero indicate little linear relationship between 2 variables.
  Even if r close to zero, there still may be a strong relationship in the form of a curve.

Interpretation of r *(opt.)*

r is a much abused statistic

- Assumption of Pearson's correlation: *at least* one variable must follow a normal distribution.
- Confidence limits are constructed for r using Fisher's z-transformation.
- r^2 is closest to 1 when n = k + 1.
  - But n should be $\geq 3(k + 1)$ for a more reliable regression model.

Significance Test for Pearson's Correlation

$H_O : \rho = 0$ (There is no linear relationship)

$H_A : \rho \neq 0$ (There is a linear relationship)

- The $H_O : \rho = 0$ is evaluated using modified t-test.
- Conclusion – significant linear correlation (i.e. $\rho \neq 0$) if p-value < 0.05

Example (Pearson)

- Correlation of cigs and weight = -0.884, p-value = 0.000 (... or rather $p < 0.001$)
- r= -0.884 suggests WHAT type of relationship?

What about if our data are only non-linearly related?

- Pearson correlation can only detect linear relationships between variables.
- Techniques are available for some non-linear relationships: Spearman's correlation coefficient can detect relationships, which are (at least) monotonic.

Cause and Effect

- Evidence of correlation does not (necessarily) mean that a cause and effect relationship exists.
- An unobserved lurking variable can be the hidden cause of an observed effect – this is referred to as spurious correlation, examples:
  - Meat consumption and cancers
  - Chocolate consumption and prostitution
  - Scotch consumption and number of teachers
  - Shoe size and vocabulary for primary school children
- What should have been measured?

Summary

- Chi-square: the analysis of frequencies
  - Applications of the χ^2: goodness-of-fit, independence, homogeneity
- Fisher's exact test: using when expected value in χ^2test statistic is small.
- Student's t test: a family of distributions, approaches the normal distribution as $n-1$ approaches infinity.
- Mann-Whitney U test: when the assumptions for using the independent t test are violated.
- Correlation: associative relationship
  - Pearson correlation can only detect linear relationships between variables.
  - For non-linear relationships: Spearman correlation coefficient.