AMS 572 Lecture Notes #10

October 28th, 2011

Ch. 9. Categorical Data Analysis

Quantitative R.V:

Numbers associated with the measurements are meaningful.

continuous R.V.: height, weight, IQ, age, etc.

discrete R.V.: time(day) , # successes, etc.

Qualitative R.V:

Numbers associated with the measurements are not meaningful.

A natural categorical variable:

Eye color	code	percentage	count
Brown	1	60%	1200
Blue	2	10%	200
Green	3	…	…
Gray	4		
Hazel	5		
Others	6		
Total		100%	2000

Sometimes we categorize quantitative data.

e.g. Age group: Children (years): <17; Young adults: [17, 35];

Middle aged adults: [36, 55]; Elderly adults: >55

1. Inference on One Population Proportion

* A special categorical R.V. -- Binary Random Variables:

Eg. Jerry has nothing to do. He decided to toss a coin 1000 times to see whether it is a fair coin. Of the 1000 tosses, he got 510 heads and 490 tails. Is a fair coin?

$$H_0 : p = \frac{1}{2}$$

$$H_a : p \neq \frac{1}{2}$$

Here the outcome variables X_i = 1 (heads) or 0 (tails), i = 0, 1, …, 1000

The total number of heads = $X_1 + X_2 + \ldots + X_{1000}$ (= 510 in this example)

* Binomial Experiment and the Binomial Distribution:

Def: A Binomial experiment consists of n trials. Each trial will result in 1of 2 possible outcomes, say "S" and "F". The probability of obtaining an "S" remains the same from trial to trial, say P. (the probability of obtaining an "F" is 1-P). These trials are independent (previous outcomes will not influence the future outcomes)

of "S" = $X \sim Bin(n, p)$ (p is population proportion)

Sample proportion: $\frac{X}{n}$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

m.g.f of X: $M_x(t) = E(e^{tx}) = \sum_{x=0}^{n} e^{tx} p(X = x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$

$$= \sum_{x=0}^{n} \binom{n}{x} (e^t p)^x (1-p)^{n-x} = (e^t p + 1 - p)^n$$

[note: $(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$, Newton's binomial theorem]

E.g. Let $X \sim Bin(n_1, p)$, $Y \sim Bin(n_2, p)$. Furthermore X and Y are independent. What is the distribution of X+Y?

Solution: $M_{X+Y}(t) = M_X(t)M_Y(t) = [e^t p + (1-p)]^{n_1+n_2}$

Hence, $X + Y \sim Bin(n_1 + n_2, p)$

When n is large. $(n \geq 30)$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n} \overset{n \text{ large}}{\sim} N\left(p, \frac{p(1-p)}{n}\right), \text{ by CLT.}$$

(Note: Here the random sample is $X_1, X_2, \ldots, X_n$: they are i.i.d. Bernoulli(p) R.V.'s.)

Bernoulli Distribution

Toss a coin, and get the result as following: Head(H), H, Tail(T), T, T, H, …

Let $X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ toss is head.} \\ 0, & \text{if the } i^{th} \text{ toss is tail.} \end{cases}$ A proportion of p is head, in the population.

$X_i \sim Bernoulli(p) \Leftrightarrow P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \quad x_i = 0, 1$

(*Binomial distribution with n = 1)

Inference on p – the population proportion:

1 Point estimator: $\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$, $\hat{p}$ is the sample proportion and also, the sample mean

2 Large sample inference on p:

$$\hat{p} \overset{\text{approx.}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

3 Large sample (the original) pivotal quantity for the inference on p.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{approx.}}{\sim} N(0,1)$$