

Soci708 – Statistics for Sociologists

Module 5 – Sampling Distributions¹

François Nielsen

University of North Carolina
Chapel Hill

Fall 2009

¹Adapted in part from slides for the course Quantitative Methods in Sociology (Sociology 6Z3) taught at McMaster University by Robert Andersen (now at University of Toronto)

Jacob Bernoulli 1st (Basel 1654–1705)

Family Were Refugees from Antwerp; *Ars Conjectandi* published 1713)



Jacob Bernoulli 1st (older)



Jacob Bernoulli 1st (stamp)

Note implausible sequence of proportions (see later)

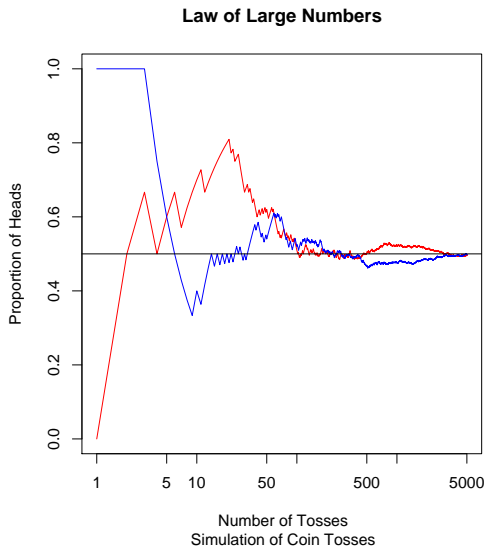


Law of Large Numbers (1)

- ▶ Jacob Bernoulli contributed to the discovery of a major phenomenon of probability, the *Law of Large Numbers*:
 - ▶ In the long run, the proportion of a certain outcome of a random trial (say, head turns up when tossing a coin) will tend to stabilize to a stable value
 - ▶ But outcome of one trial is *independent of previous outcomes*
- ▶ This is *counterintuitive*:
 - ▶ People naturally tend to believe in a sort of *Law of Small Numbers*
 - ▶ People do not normally expect the long “runs” of the same outcome (say, heads in tossing a coin) that occur in true random processes

Law of Large Numbers (2)

Two simulations of tossing a fair coin 5,000 times



Sampling Distributions Revisited

Population & Sample

- ▶ The *population distribution* of a variable is:
 1. the distribution of its values for all members of the population
 - ▶ E.g., the distribution of IQ test scores in the Belgian population
 2. the probability distribution of the variable when choosing one individual at random from the population.
 - ▶ E.g., choose one Belgian randomly and record the IQ
- ▶ A *statistic* (e.g., \bar{x} , \hat{p} , b_1) calculated from a random sample or randomized experimental group is a random variable
- ▶ The probability distribution of a statistic is its *sampling distribution*
- ▶ In remainder of Module 6 we look at the sampling distributions of:
 - ▶ counts & proportions
 - ▶ sample means

Binomial Distributions

Count X & Proportion \hat{p}

- ▶ In general X is a *count* of the occurrence of some outcome in a fixed number of observations n
 - ▶ E.g., in an agricultural experiment n plants are treated for a fungus; the number X of plants with the fungus is a random variable
- ▶ The *sample proportion* is $\hat{p} = X/n$
 - ▶ E.g., in the experiment $X = 9$ out of $n = 32$ plants have the fungus. The sample proportion is

$$\hat{p} = \frac{9}{32} = 0.281$$

- ▶ The *binomial setting* is:
 1. There are a fixed number n of observations
 2. The n observations are all independent
 3. Each observation can be classified as “success” (1) or “failure” (0)
 4. The probability p of a success is the same for each observation

Binomial Distributions

Binomial Distribution

- ▶ The distribution of the count X of successes in the binomial setting is called the *binomial distribution* with parameters n (number of observations) and p (probability that any one observation is a success)
 - ▶ The possible values of X are the positive integers from 0 to n
 - ▶ In abbreviation, one says that X is $B(n, p)$
- ▶ E.g., a child of a specific couple has probability $p = 0.25$ of being blood type O. Suppose the couple has $n = 5$ children. Then the number X of their children with blood type O is distributed as $B(5, 0.25)$
 - ▶ Possible values of X are 0, 1, 2, 3, 4, 5
 - ▶ The probability distribution of X is (see why later)

$X :$	0	1	2	3	4	5
$P(X = x) :$	0.2373	0.3955	0.2637	0.0879	0.0146	0.001

Binomial Distributions

Binomial Distribution

- ▶ Choosing an SRS (without replacement) from a population with proportion p of successes is *not exactly* a binomial setting
 - ▶ E.g., draw 10 cards from a deck, with “red card” a success. Then probability of red on second card is not independent of color of first card
- ▶ However, if the population is much larger than the sample – say, 20 times as large – the count X of successes in an SRS of size n has approximately the binomial distribution $B(n, p)$
 - ▶ E.g., draw sample with $n = 200$ from about 8,000 graduate students at UNC. “Success” is: student is female. Suppose $p = 0.57$. Then number of females X is distributed (almost exactly) as $B(200, 0.57)$

Binomial Distributions

Finding Binomial Probabilities (1)

1. Calculator on the Web

- ▶ <http://rockem.stat.sc.edu/prototype/calculators/index.php3>

2. Table of Binomial Probabilities

- ▶ E.g., Table C in Moore & McCabe (2006)

3. Software – R

- ▶ Finding $P(X = x)$

```
> # P(exactly 2 children out of 5 with type 0 blood)
> dbinom(2, 5, 0.25)
[1] 0.2636719
```

- ▶ Finding $P(X \leq x)$

```
> # P(2 or fewer children out of 5 with type 0 blood)
> pbinom(2, 5, 0.25)
[1] 0.8964844
```

Binomial Distributions

Finding Binomial Probabilities (2)

4. Software – Stata

- ▶ Finding $P(X = x)$
 - . * P(exactly 2 children out of 5 with type 0 blood)
 - . display Binomial(5,2,0.25) - Binomial(5,3,0.25)
 - .26367188
- ▶ Finding $P(X \leq x)$
 - . * P(2 or fewer children out of 5 with type 0 blood)
 - . display 1 - Binomial(5,3,0.25)
 - .89648438
- ▶ *Note:* In Stata the function `Binomial(n,k,p)` returns $P(X \geq x)$. It has to be spelled with capital B.

Binomial Distributions

Finding Binomial Probabilities (3)

5. Using the Binomial Formulas (Optional; see Moore & McCabe 2006, pp.348–350)

- ▶ Binomial Coefficient – The number of ways of arranging k successes among n observations is given by the *binomial coefficient*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, \dots, n$. In the formula the factorial $n!$ for any positive integer is defined as

$$n! = n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$$

and also $0! = 1$.

- ▶ Binomial Probability – If X has distribution $B(n, p)$, the *binomial probability* that $X = k$ (for $k = 0, 1, \dots, n$) is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distributions

Origin of the Binomial Formula

- ▶ Origin of the binomial formula

Binomial Distributions

Binomial Mean & Standard Deviation

- ▶ If a count X is $B(n, p)$, what are the mean μ_X and the standard deviation σ_X of X ?
- ▶ To find out view X as the sum of n independent random variables S_i . Each S_i has the same probability distribution

Outcome:	1	0
Probability:	p	$1 - p$

- ▶ For a single S_i (which, BTW, is called a *Bernoulli trial*)

$$\mu_S = (1)(p) + (0)(1 - p) = p$$

$$\sigma_S^2 = p(1 - p)$$

- ▶ Then for $X = S_1 + S_2 + \cdots + S_n$

$$\mu_X = \mu_{S_1} + \mu_{S_2} + \cdots + \mu_{S_n} = n\mu_S = np$$

$$\sigma_X^2 = n\sigma_S^2 = np(1 - p)$$

Binomial Distributions

Mean & Standard Deviation of Count & Proportion

- ▶ If a count X has the binomial distribution $B(n, p)$, then

$$\mu_X = np$$

$$\sigma_X = \sqrt{np(1-p)}$$


- ▶ Our estimator of the proportion p of “successes” in the population is the sample proportion

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

- ▶ If \hat{p} is the sample proportion of successes in an SRS of size n from a large population with proportion p of successes²

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

²Check this follows from the rules for linear functions of random variables. 

Binomial Distribution

Normal Approximation of Counts & Proportions

- ▶ Implications of mean and standard deviation of \hat{p}
 1. $\mu_{\hat{p}} = p$ implies \hat{p} is *unbiased*
 2. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ implies that to divide the standard deviation of \hat{p} by half one must multiply n by four
- ▶ *Normal approximation for counts & proportions:*
 - ▶ In an SRS of size n from a large population, when n is large

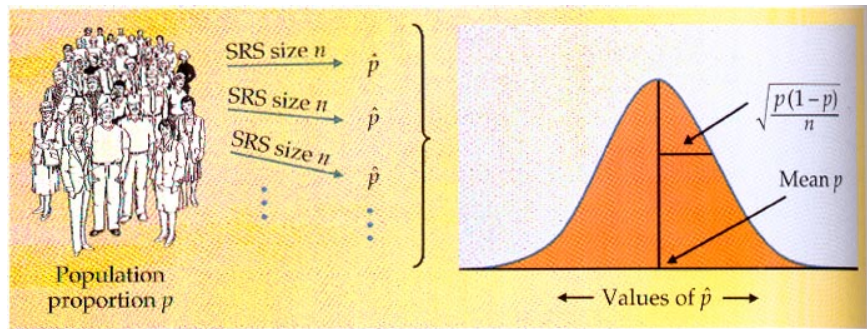
X is approximately $N\left(np, \sqrt{np(1-p)}\right)$

\hat{p} is approximately $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

where p is the proportion of successes in the population, and X and $\hat{p} = X/n$ are the count & proportion of successes in the sample, respectively

Binomial Distribution

Normal Approximation of Counts & Proportions



- ▶ Rule of thumb for normal approximation: n & p satisfy
 - ▶ $np \geq 10$, and
 - ▶ $n(1-p) \geq 10$

Binomial Distribution

Normal Approximation of Counts & Proportions

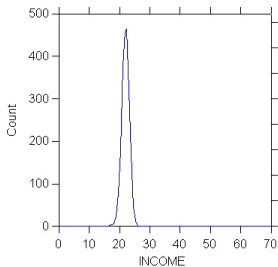
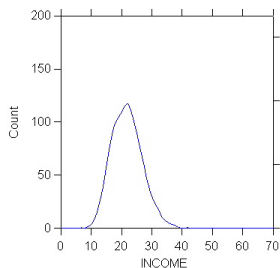
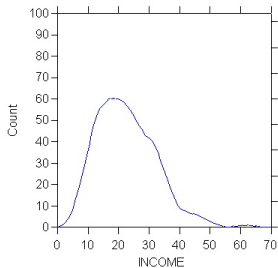
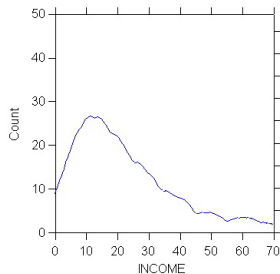
- ▶ E.g., SRS of $n = 200$ from population of 8,000 UNC graduate students with proportion females $p = .57$. What is $P(\hat{p} \leq 0.5)$ (i.e., the sample has fewer females than males)?
 - ▶ $np = 200 \times 0.57 = 114 > 10$ and $n(1 - p) = 200 \times 0.43 = 86 > 10$ so rule of thumb is satisfied
 - ▶ Using binomial probabilities: X is distributed as $B(200, 0.57)$. $\hat{p} = 0.5$ correspond to $X = 100$. $P(X \leq 100) = 0.02734091$ or .027.
 - ▶ Using the normal approximation: $\mu_{\hat{p}} = p = 0.57$;

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.03500714;$$

$$\begin{aligned} P(\hat{p} \leq 0.5) &= P\left(\frac{\hat{p} - 0.57}{0.03500714} \leq \frac{0.5 - 0.57}{0.03500714}\right) \\ &= P(Z \leq -1.999592) = 0.02277217 \end{aligned}$$

Sampling Distribution of the Sample mean

Experimental Study of the Sampling Distribution of \bar{x} with $n = 3, n = 10, n = 100$



- ▶ (a) Population distribution of X (income)
- ▶ Distribution of \bar{x} for 600 samples:
 - ▶ (b) $n = 3$
 - ▶ (c) $n = 10$
 - ▶ (d) $n = 100$

Sampling Distribution of the Sample mean

Experimental Study of the Sampling Distribution of \bar{x} with $n = 3$, $n = 10$, $n = 100$

Income Sampling Experiment

Data	Mean	SD
Population	22.172	15.635
$n = 3$	22.584	9.376
$n = 10$	21.955	4.916
$n = 100$	22.176	1.193

- The experimental results suggest the following conjectures:
1. The distribution of values of \bar{x} for a SRS is centered around the population mean μ_X , regardless of sample size
 2. The standard deviation $\sigma_{\bar{x}}$ of values of \bar{x} decreases with increasing sample size – i.e., as n increases the distribution of \bar{x} values becomes more concentrated around the population mean μ_X
 3. The distribution of \bar{x} values becomes more symmetrical as the sample size becomes larger and is approximately normal for large n s

Sampling Distribution of the Sample mean

Theoretical Development: Mean & Standard Deviation of \bar{x}

- ▶ The mean \bar{x} of a SRS is a random variable

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

- ▶ If the population has mean μ then by the addition rule for a sum of random variables

$$\begin{aligned}\mu_{\bar{x}} &= \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n}) \\ &= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu\end{aligned}$$

- ▶ Thus *the mean of \bar{x} is μ , the same as the mean of the population*
 - ▶ I.e., \bar{x} is an *unbiased* estimator of μ

Sampling Distribution of the Sample mean

Theoretical Development: Mean & Standard Deviation of \bar{x}

- ▶ Because the observations X_i are independent, by the addition rule for variances

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \left(\frac{1}{n}\right)^2 (\sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_n}^2) \\ &= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

- ▶ Thus for an SRS of size n from population with mean μ and standard deviation σ

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Sampling Distribution of the Sample mean

Experimental Study of the Sampling Distribution of \bar{x} with $n = 3$, $n = 10$, $n = 100$

Table 21. Income Sampling Experiment Results and Theoretical Values Compared (600 Samples)

Data	Mean	SD	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}$	Fpcf ³	$\sigma_{\bar{x}}^{*4}$
Population	22.172	15.635	—	—	—	—
$n = 3$	22.584	9.376	22.172	9.027	0.994	8.974
$n = 10$	21.955	4.916	22.172	4.944	0.980	4.846
$n = 100$	22.176	1.193	22.172	1.564	0.781	1.221

³Finite population correction factor $\sqrt{1 - \frac{n}{N}}$ with $N = 256$ and $n = 3, 10, 100$

⁴Finite population corrected standard error $\sigma_{\bar{x}}^* = \sigma_X / \sqrt{n} \times \sqrt{1 - \frac{n}{N}}$

Sampling Distribution of the Sample mean

Why Does the Distribution of \bar{x} Become Normal When n Increases?

- ▶ In income sampling experiment:
 - ▶ The distribution of income in the population is *not* normal (it is skewed to the right)
 - ▶ *Even so*, the distribution of sample mean \bar{x} becomes symmetric & “normal-looking” when n increases
- ▶ This is due to a very important *natural phenomenon*, called the *Central Limit Theorem*:
 - ▶ Draw an SRS of size n from *any* population with mean μ and finite standard deviation σ . When n is large, the sampling distribution of the sample mean \bar{x} is approximately normal, so that

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- ▶ The normal approximation for sample proportions & counts is also an instance of the CLT
- ▶ Special case: the mean of an SRS from a normal population is also normally distributed (for any n)