Statistics I – Chapter 7
Sampling Distributions (Part 2)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

November 21, 2012

Road map

- **Distribution of the sample proportion**.
- Correction for finite populations.
- Distribution of the sample variance.
- Proof of the central limit theorem.

Means vs. proportions

- For interval or ratio data, we have defined sample means.
  - We have studied the distributions of sample means.
- For **ordinal or nominal** data, there is no sample mean.
  - Instead, there are sample **proportions**.

Population proportions

- How to know the **proportions** of girls and boys in NTU?
- We first **label** girls as 0 and boys as 1.
- Let $X_i \in \{0, 1\}$ be the sex of student i, $i = 1, ..., N$.
- Then the **population proportion** of boys is defined as

$$p = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- The population proportion of girls is $1 - p$.

Sample proportions

- Let $\{X_i\}_{i=1,\ldots,N}$ be the population.
- With a sample size n, let $\{X_i\}_{i=1,\ldots,n}$ be a sample. Suppose X_i and X_j are independent for all $i \neq j$.
  - E.g., 100 randomly selected students.
- Then the **sample proportion** is defined as

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- The population proportion p is deterministic (though unknown) while the sample proportion $\hat{p}$ is **random**.
- We are interested in the distribution of $\hat{p}$.

Examples of sample proportions

- Proportion of voters preferring a particular candidate.
- Proportion of employees in the manufacturing industry.
- Proportion of faculty members hired in six years.
- Proportion of people higher than 180 cm.

Distributions of sample proportions

- What is the distribution of the sample proportion

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i?$$

- As X_i is the outcome of a randomly selected entity, it follows the population distribution.
  - Therefore, $X_i \sim \text{Ber}(p)$.
- It then follows that $\sum_{i=1}^{n} X_i \sim \text{Bi}(n, p)$.
- But is $\frac{1}{n}\sum_{i=1}^{n} X_i$ also binomial?

Distributions of sample proportions

- Let $X_1 \sim \text{Bi}(n_1, p)$ and $X_2 \sim \text{Bi}(n_2, p)$ where X_1 and X_2 are independent. Consider $\frac{1}{2}(X_1 + X_2)$.
- Can it follow a binomial distribution?
- No! Why?
- Then what may we do?

Distributions of sample proportions

- One thing we have learned is to use a **normal** distribution to approximate a binomial distribution.
  - If $n \geq 25$, $np < 5$, and $n(1-p) < 5$, we have

$$\sum_{i=1}^{n} X_i \sim \text{ND}\Big(np, \sqrt{np(1-p)}\Big).$$

  - So $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \text{ND}(p, \sqrt{\frac{p(1-p)}{n}})$.
- Or we may apply the **central limit theorem**:
  - If $n \geq 30$, a sample mean ($\hat{p}$ in this case) is approximately normally distributed:

$$\mathbb{E}[\hat{p}] = \mu = p \quad \text{and} \quad \text{Var}(\hat{p}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}.$$

- If n is small, we need to derive the distribution by ourselves.

Sample proportions: An example

- In 2011, there are 19756 boys and 13324 girls in NTU.
- The population proportion of boys is

$$p = \frac{19756}{33080} \approx 0.597.$$

- Suppose we sample 100 students and calculate the sample proportion $\hat{p}$.
  - What is the distribution of $\hat{p}$?
  - What is the probability that in the sample there are fewer boys than girls?

Sample proportions: An example

- What is the distribution of $\hat{p}$?
  - As $n \geq 30$, it follows a normal distribution.
  - Its mean is $p \approx 0.597$.
  - Its standard deviation is $\sqrt{\frac{p(1-p)}{n}} \approx 0.049$.
- What is the probability that $\hat{p} < 0.5$?

$$\Pr(\hat{p} < 0.5) = \Pr\left(Z < \frac{0.5 - 0.597}{0.049}\right)$$
$$\approx \Pr(Z < -1.98) \approx 0.024.$$

Sample proportions: Remarks

- A sample proportion "is" a sample mean of qualitative data.
- It is normal when the sample size is large enough.
  - A binomial distribution approaches a normal distribution.
  - A sample mean approaches a normal distribution.
- In using statistics to estimate parameters:
  - We use a sample proportion $\hat{p}$ to estimate the population proportion p.
  - We use a sample mean $\overline{X}$ to estimate the population mean μ.
- It is intuitive, but is it good?
- We will study this in Chapter 8.

Road map

- Distribution of the sample proportion.
- **Correction for finite populations**.
- Distribution of the sample variance.
- Proof of the central limit theorem.

Sample means revisited

- For the sample mean and sample proportion, the sample should be **independent**.
  - $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. X_i and X_j are independent for all $i \neq j$.
- What if they are not independent?
  - Is the variance still $\frac{\sigma^2}{n}$ or $\frac{p(1-p)}{n}$?
  - Is the sample mean still normal with a normal population?
  - Is the sample sum still binomial with a Bernoulli population?
  - Does the central limit theorem still hold?

Sample means revisited

- Most of the sampling in practice are **sampling without replacement**.
- Only if the population size is **large** enough (compared with the sample size), samples generated by sampling without replacement can be treated as independent.
  - A rule of thumb is $n < 0.05N$.
- When the population size is not large enough, we say we sample from a **finite population**.
- What should we do in this case?

Finite populations: variances?

- Question 1: Is the variance still $\frac{\sigma^2}{n}$ or $\frac{p(1-p)}{n}$?
- When sampling from a finite population, we may fix the variance of the sample mean.
- Recall that for $X \sim \mathrm{HG}(N, A, n)$, we have

$$\mathrm{Var}(X) = np(1-p)\left(\frac{N-n}{N-1}\right), \quad \text{where } p = \frac{A}{N}.$$

- The coefficient $\frac{N-n}{N-1}$ is called the **finite correction factor** of **variance**.
- $\sqrt{\frac{N-n}{N-1}}$ is the finite correction factor of **standard deviation**.

Finite populations: variances?

- It can be shown that, when sampling from a finite population, the **sample mean's variance** should also contain the finite correction factor:

$$\mathrm{Var}(\overline{X}) = \left(\frac{\sigma^2}{n}\right)\left(\frac{N-n}{N-1}\right).$$

  - The derivation is similar to what we have done in homework.

Finite populations: normal?

- Question 2: Is the sample mean still normal when the population is normal?
- If we sample from a **normal population**, the sample mean is normal even if the sample is not independent.
  - Sum of two (or n) **dependent** normal random variables is still normal.

Finite populations: binomial?

- Question 3: Is the sample sum still binomial when the population is Bernoulli?
- For qualitative populations, we know if the population size is large, the sample sum follows a **binomial** distribution.
- If the population size is small, the sample sum follows a **hypergeometric** distribution.
- The distribution of sample proportion can then be determined (though the calculation is quite tedious).
- When it is impossible to derive the distribution of sample proportion, use approximations.

Finite populations: CLT?

- Question 4: Does the central limit theorem hold?
- The central limit theorem we learned in the last lecture does require independence.
- Without independence, there are **generalized** versions of the central limit theorem.
  - We may still have normality when we lose independence.
  - We will not touch these generalized versions.
- Nevertheless, we will still "pretend" that the usual central limit theorem applies and assume the sample mean and sample proportion are normally distributed.

Finite populations: conclusions

- If we sample from a finite population (i.e., $n > 0.05N$):
  - If $n \geq 30$, we will still assume the sample mean and sample proportion are normally distributed.
  - Their variances will be multiplied by $\frac{N-n}{N-1}$.
  - If $n < 30$, we need to derive the sampling distributions for the two statistics by ourselves.

Road map

- Distribution of the sample proportion.
- Correction for finite populations.
- **Distribution of the sample variance**.
- Proof of the central limit theorem.

Sample variances

- Let $\{X_i\}_{i=1,\ldots,n}$ be a random sample. The **sample variance** is defined as
$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$
- The **sample standard deviation** is $S = \sqrt{S^2}$.
- A sample variance is **not** the variance of a sample mean!
  - The sample variance is a random variable.
  - The variance of the sample mean is a fixed number.

Sample variances

- As a sample variance is random, it has its own distribution.
- While it is hard to derive the distribution of S^2, it is easier (though still not very easy) to derive the distribution of

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}.$$

- Why do we care about the distribution of this statistic?
  - If we know the distribution of this statistics χ^2, we will be able to infer σ^2 from the realization of S^2.
  - This will be discussed in Chapter 8.

Distribution of sample variances

- As the statistic is called χ^2, you probably can guess what is the sampling distribution:

Proposition 1

Let $\{X_i\}_{i=1,\ldots,n}$ *be a random sample from a normal population with variance* σ^2, *i.e.,* $\mathrm{Var}(X_i) = \sigma^2$. *Then*

$$\chi^2 \equiv \frac{(n-1)S^2}{\sigma^2}$$

follows the chi-square distribution with degree of freedom $n-1$, *i.e.,* $\chi^2 \sim \mathrm{Chi}(n-1)$.

Chi-square distributions

- We have defined the **chi-square (χ^2) distribution** in Chapter 6:

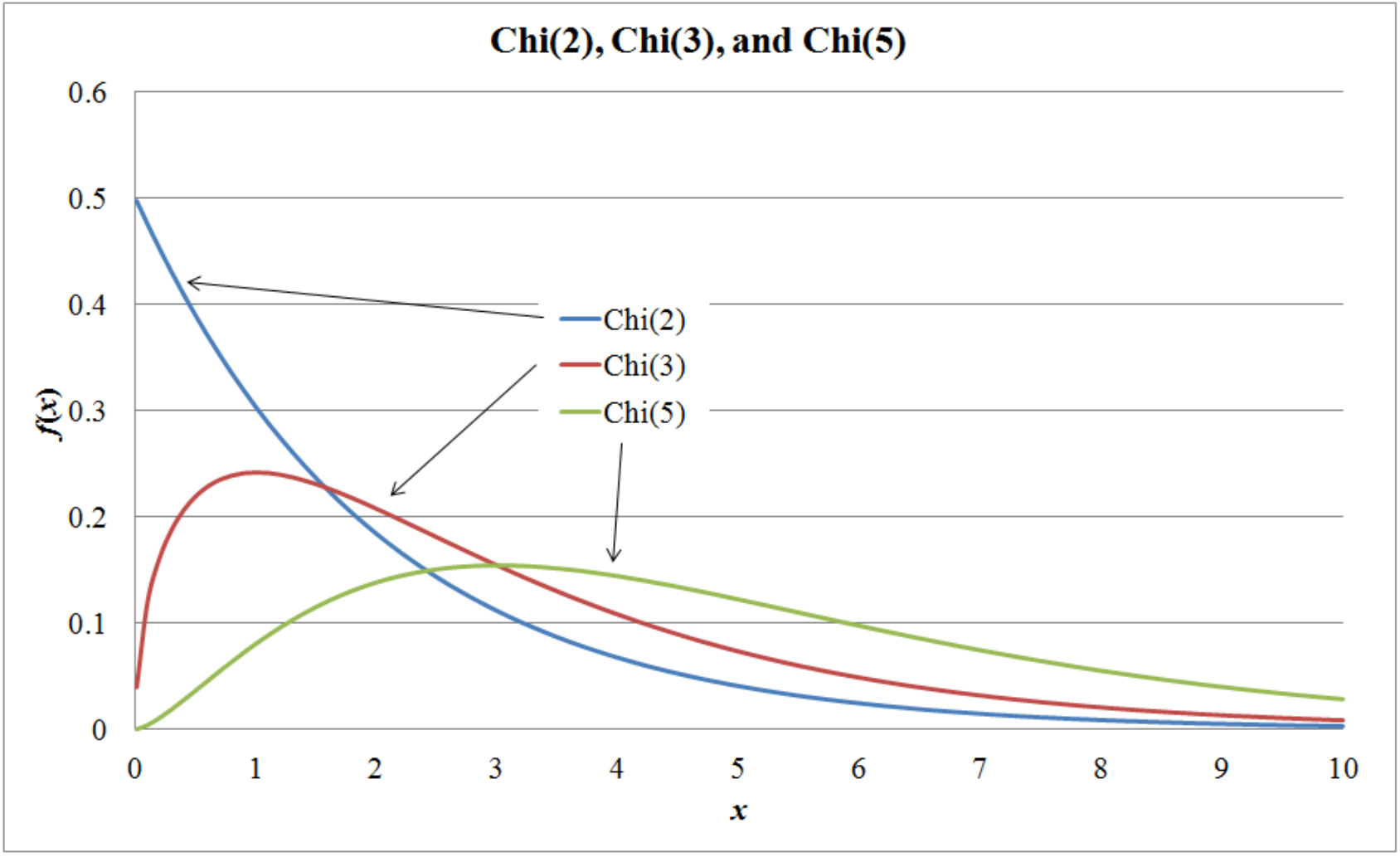
Definition 1 (Chi-square distribution)

A random variable X follows the chi-square distribution with degree of freedom $n \in N$, denoted by $X \sim \chi^2(n)$ or $X \sim \text{Chi}(n)$ if it follows the gamma distribution with $\alpha = \frac{n}{2}$ and $\beta = 2$.

- With $\Gamma(x) = \int_0^\infty e^{-x} x^{z-1} dx$, the pdf of $X \sim \text{Chi}(n)$ is

$$f(x|n) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{\sqrt{2^n}\Gamma(\frac{n}{2})} \quad \forall x \geq 0.$$

Chi-square distributions



Proof of the proposition

- Let's now prove the proposition with markers and the white board.

Road map

- Distribution of the sample proportion.
- Correction for finite populations.
- Distribution of the sample variance.
- **Proof of the central limit theorem**.

Proof of the central limit theorem

- Let's now prove the central limit theorem with markers and the white board.