Introduction to Statistics
Lecture 16

Reminder: Quiz 3 next week.

Binomial model (cont'd)

- For n random Bernoulli trials, X is the number of successes. If the probability of success is p, q=1-p is the probability of failure,

"n choose k"

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

$$5! = 5\times4\times3\times2\times1$$

$$0! = 1$$

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

Mean: ***np***

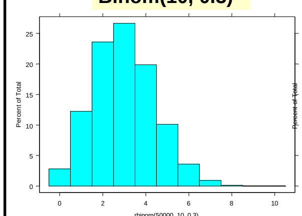
Standard deviation: $\sqrt{np(1-p)}$

Example

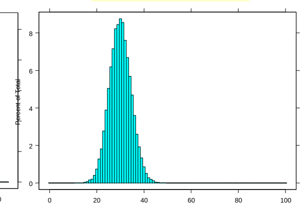
- A large pool of candies. 30% red.
- We randomly sampled 10, what is the probability 5 of them are red?
- What is the probability that fewer than 5 of them are red?

Normal approximation of binomial distribution

Binom(10, 0.3)



Binom(100, 0.3)



When n is large, so that (np>10) and (nq>10), the binomial distribution Binom(n, p) can be approximated by $N(np, \sqrt{np(1-p)})$

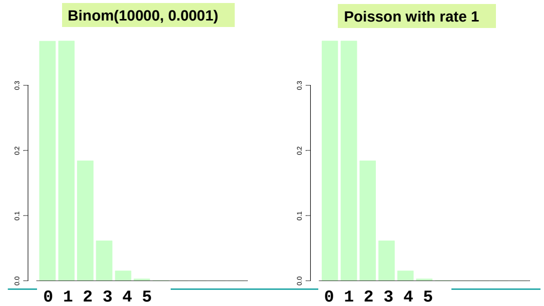
Calculation steps
Normal approximation of Binomial model

- Check Bernoulli trials?
  - Success/failure
  - Probability of success
  - Independence
- Find **n**—number of trials
- Find **p**—probability of success for each trial
- Write down the distribution **Binom(n, p)**
- Check Normal approximation rules:
  - np>10
  - n(1-p)>10
- Find mean and standard deviation of Binom(n, p)
- Use normal distribution with the same mean and standard deviation to carry out the calculation.

What if p is too small that np<10 even when n is large

- Poisson distribution can be used to approximate the binomial distribution when
  - n is large
  - np is small
- where λ=np. $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$

Poisson approximation of the binomial distribution



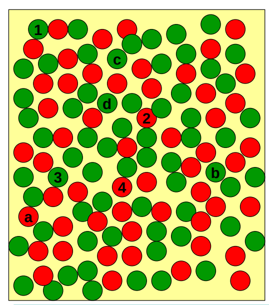
Why study binomial distribution and its approximations

- Consider the following scenario:
  - In a large population, 60% of the people support the current mayor.
  - You random sample one individual
    - Success: he/she supports the mayor
    - Failure: he/she doesn't support the mayor
    - Probability of success = 60% ?
  - A survey of 100 people, X is the number of people that supports the mayor in this survey.

Why probability models are important? Or why binomial distribution is important?

We are starting chapters on statistical inference NOW.

Different samples out of a population



	Sample 1 (1,2,3,4)	Sample 2 (a,b,c,d)
1	Green	Red
2	Red	Green
3	Green	Green
4	Red	Green

Sample proportion

- Given a simple random sample with n observations of a Bernoulli trial
- X: the number of success
- p: probability of success
- Sample proportion: $$\hat{p} = \frac{X}{n}$$
- **Sample proportion is a random variable** since X is a random variable that has Binomial model.

Sampling distribution models of the sample proportion

- **Sampling variation**: the p-hat calculated based on different random samples from the sample population differs due to chance.
- Parameters:
  - The sample size, **n**
  - The probability of success: long-term relative frequency (proportion) of success in the population—**population proportion, p**
- We can study the variation of sample proportion using some probability model under some assumptions—**sampling distribution model.**

By the normal approximation …

X is approximately normally distributed with mean np and standard deviation $\sqrt{np(1-p)}$

Then,

$\hat{p} = \dfrac{X}{n}$ is approximately normally distributed with mean p and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$.

Sampling distribution model of sample proportion!

General situation …

- N: population size
- n: sample size
- To use normal approximation of binomial model to study the random behavior of the sample proportion, we need
  - n < N/10
  - np > 10 AND nq=n(1-p) > 10

Example

- As historically studied,15% of faculty members leave campus during spring break. For a (simple) random sample of 100 faculty members, what is the chance that more than 20% of them left campus during the past spring break?

We just discussed about proportion, what about mean, the average?

Let's see an online demo first.

Mean and Variance of Sample mean

n observations: $X_1, \ldots, X_n$

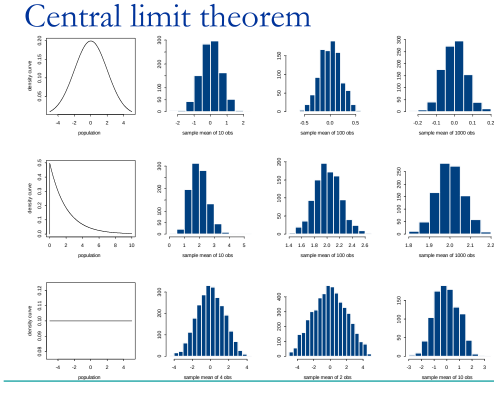
independent, and have the same probability distribution

$$\mu_{X_1} = \mu \quad \sigma^2_{X_1} = \sigma^2$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ so } \mu_{\bar{X}} = \frac{1}{n}(\mu + \ldots + \mu) = \mu$$

$$\sigma^2_{\bar{X}} = \left(\frac{1}{n}\right)^2 \left(\sigma^2 + \ldots + \sigma^2\right) = \frac{\sigma^2}{n}$$

Central limit theorem



Sampling distribution model of Sample mean

- If the population distribution is $N(\mu, \sigma)$

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- If the population distribution is not normal and with mean μ and standard deviation σ

$$\boxed{\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \text{ approximately when n is large.}}$$

- How large is large?

In practice

- Standard error: estimated standard deviation of a sampling distribution.
- Distinguish the sampling distribution and the distribution of a sample.

Reading

- Chapter 18