Lec 5*: Sampling Distributions

1 Motivation

- Last lecture, we discussed random variables, and the probability distributions that such random variables could follow.
- As we learned from last lecture,
  - A random variable assign a number to each possible outcome.
  - A discrete probability distribution describes the point probability at all possible values for a discrete random variable.
  - A continuous probability distribution describes the density (PDF) at all possible values for a continuous random variable.

  Thus, random variables and their associated probability distributions are related to the **population**.
- However, in reality, what we get to work with is often the **sample** data, which means we need to relate statistics obtained from samples to the population ($\Rightarrow$ process of statistical inference).
- This is why we need to look at the distribution of sample statistics, i.e. **sampling distributions**

2 Difference Between Probability Distribution and Sampling Distribution

	Probability Distribution	**Sampling Distribution**
Generated by ...	Random variable (e.g. X)	Sample statistic (e.g. $\bar{X}$)
Describes ...	Probability of a random variable equals to a certain value	Probability of a sample statistic equals to a certain value
Helps us know about ...	How likely a number is drawn from the population	How likely the sample statistic is calculated as some number

3 Examples of Sampling Distribution

3.1 Sampling distribution of the mean

- Statistic of interest: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, obtained from simple random sampling
- How $\bar{X}$ is distributed depends on the distribution of X_i:
  - If each X_i is normally distributed, then $\bar{X} \sim N(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$ with certainty
  - If each X_i is NOT normally distributed, we might be able to approximate $\bar{X}$ using a normal distribution (i.e. $\bar{X} \overset{a}{\sim} N(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$) based on central limit theorem.

*Some exercise questions are taken from or slightly modified based on Dr. Gregory Pac's Econ 310 discussion handout.

Theorem 1 (Central limit theorem (CLT)). The mean of a random variable drawn from any population is approximately normal for a sufficiently large sample size.

In practice, we use $n \geq 30$ as the cutoff:

* For non-normally distributed X_i, if $n \geq 30$, then CLT can be invoked, and $\bar{X} \overset{a}{\sim} N(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$
* For non-normally distributed X_i, if $n < 30$, then CLT cannot be invoked, so the distribution of $\bar{X}$ is undetermined.

- To summarize, for random variable X, **the sampling distribution of the mean is the following**:

	X **is normally distributed**	X **is NOT normally distributed**
Sample size is small ($n < 30$**)**		
Sample size is large ($n \geq 30$**)**		

- What are $\mu_{\bar{X}}$ and $\sigma^2_{\bar{X}}$?
  - $\mu_{\bar{X}}$ is the expected value of $\bar{X}$:

$$\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n \times \mu_X = \mu_X$$

  - $\sigma^2_{\bar{X}}$ is the variance of $\bar{X}$, and it depends on the population size:
    * If population size is infinitely large (in practice, if $N \geq 20n$),

$$\sigma^2_{\bar{X}} = V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} \times n \times \sigma^2_X = \frac{\sigma^2_X}{n}$$

    * If population size is not infinitely large (in practice, if $N < 20n$), then $\sigma^2_{\bar{X}}$ needs to be adjusted:
      · **Finite population correction factor**: an adjustment applied to the **standard error** of sample mean (i.e. $\sigma_{\bar{X}}$), where

$$\text{Finite population correction factor} = \sqrt{\frac{N-n}{N-1}}$$

      · Thus, the standard error of sample mean is

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2_X}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

      which means that the variance of the sample mean is

$$\sigma^2_{\bar{X}} = (\sigma_{\bar{X}})^2 = \frac{\sigma^2_X}{n} \cdot \frac{N-n}{N-1}$$

3.2 Sampling distribution of the proportion (from a binomial experiment)

- Say that we have a random variable $X \sim \text{Binomial}(n, p)$ recording the number of successes in n trials where the probability of success in each trial is p.
- Turns out, under certain conditions, X can be well approximated by a normal distribution.

> Conditions for normal approximation of a binomial random variable X:
>
> 1. $np \geq 5$, and
> 2. $n(1-p) \geq 5$

If the aforementioned conditions are satisfied, then

$$X \overset{a}{\sim} N(\mu_X, \sigma_X^2)$$

where, based on binomial distribution properties,

$$\mu_X = E(X) = np$$
$$\sigma_X^2 = V(X) = np(1-p)$$

> Aside: A binomial X is a discrete random variable. However, the approximation approximates $X \overset{a}{\sim} N(np, np(1-p))$, which is a continuous distribution.
>
> Thus, a **correction factor for continuity** is needed when calculating probability using the normal approximation.
> Exercise. Accounting for the correction factor for continuity, how should the following probabilities be expressed for a binomial random variable X?
>
> 1. $P(X = 3) =$
> 2. $P(X = 2) =$
> 3. $P(X = 2 \text{ or } 3) =$

- Why is this needed? $\Rightarrow$ helps us approximate the sampling distribution of the proportion!
  - As long as a binomial distributed X can be approximated using a normal distribution (i.e. $np \geq 5$ and $n(1-p) \geq 5$), then the proportion of successes ($\hat{p}$) can be approximated using a normal distribution:

$$\hat{p} = \frac{X}{n} \overset{a}{\sim} N(\mu_{\hat{p}}, \sigma_{\hat{p}}^2)$$

  - What is $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}^2$?

$$\mu_{\hat{p}} = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = p$$

$$\sigma^2_{\hat{p}} = V(\hat{p}) = V\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 V(X) = \frac{p(1-p)}{n}$$

3.3 Sampling distribution of the difference between two means

- Statistic of interest: $\bar{X} - \bar{Y}$, where $X \sim N(\mu_X, \sigma^2_X)$, and $Y \sim N(\mu_Y, \sigma^2_Y)$, and X is independent of Y
- From subsection 3.1, assuming that the population sizes are sufficiently large, we know that

$$\bar{X} \sim N(\mu_X, \frac{\sigma^2_X}{n_X}) \qquad \bar{Y} \sim N(\mu_Y, \frac{\sigma^2_Y}{n_Y})$$

- Since the sum of two normal distributions is still a normal distribution, we have

$$\bar{X} - \bar{Y} \sim N(\mu_{\bar{X}-\bar{Y}}, \sigma^2_{\bar{X}-\bar{Y}})$$

where

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$
$$\sigma^2_{\bar{X}-\bar{Y}} = V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) - 2\underbrace{Cov(\bar{X}, \bar{Y})}_{=0 \text{ by indep}} = \frac{\sigma^2_X}{n_X} + \frac{\sigma^2_Y}{n_Y}$$

4 Exercises

1. Suppose we draw a simple random sample of four observations: $\{X_1, X_2, X_3, X_4\}$. Each X_i is distributed with mean 4 and standard deviation 2. The realized values for our sample turn out to be: $\{-1, 0, 5, 3\}$.
   (a) What is $E(\bar{X})$? Would your answer change if you were working with a different sample, such as: $\{4, -1, 2, 6\}$?

(b) What is $V(\bar{X})$? Would your answer change if you were working with a different sample?

(c) What is the distribution of $\bar{X}$?

(d) Now suppose $n = 64$. What is the distribution of $\bar{X}$?

(e) Now suppose $n = 64$, and $X_i \sim N(4, 4)$. What is the distribution of $\bar{X}$?

2. The amount of time a bank teller spends with each customer has a population mean $\mu = 3.1$ minutes and a standard deviation of $\sigma = 0.4$ minutes.

   (a) If a random sample of 50 customers is selected from a finite population of 500 customers, what is the probability that the average time per customer will be at least 3 minutes?

   (b) Now, suppose that we observe only 16 customers, and answer the same question.

3. Let X be the number of successes in a binomial experiment with $n = 300$ and $p = 0.55$, and let $\hat{p} = \frac{X}{n}$ be the proportion of successes.

   (a) Is this a case where X is well approximated by a normal distribution? If so, exactly what normal distribution should we use?

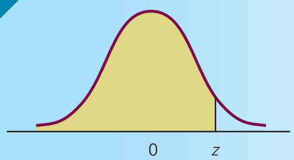
   (b) Using a normal approximation, what is the probability that $X = 165$? Use the correction factor for continuity.

(c) Is this a case where $\hat{p}$ is well approximated by a normal distribution? If so, exactly what normal distribution should we use?

(d) Find the approximate probability that $\hat{p}$ is greater than 60%.

(e) We would like to repeat the same binomial experiment with $p = 0.55$, but with fewer trials. If we want to use the normal distribution to approximate $\hat{p}$, how many trials do we need?

TABLE **3** ***(Continued)***



$P(-\infty < Z < z)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990