Urban Design and Planning
URBDP 520 Quantitative Methods in Urban Design and Planning

Lecture Notes 5: Sampling Distributions

Introduction

There are two really interesting things to do in statistics.

The first is to be able to predict what a sample drawn from a population is likely to look like, given what you know about the population and the size of the sample.

The second is to be able to predict what a population is likely to look like, given what you know about a sample which has been drawn from it.

In this chapter, we'll investigate the first of these. While there are some situations in which the first thing is very compelling, the primary value of the answers we'll find is in turning the question around and approaching the second of these.

Sampling...

Sampling is as much an art as it is a science. Drawing a truly random sample from a population can be very tricky and has only recently been practiced at a level nearing perfection.

Populations can be regarded as finite or infinite. A finite population can be listed (the members of a class or the citizens of the U.S.) while an infinite population cannot be listed (all possible customers at McDonald's in a day).

In the case that a population is finite, the text *How to Conduct Your Own Survey* by Priscilla Salant and Don Dillman offers the following three steps in sampling (p. 58)

1. Identify the target population as precisely as possible and in a way that makes sense in terms of the purpose of the study. It is important to be specific enough that everyone involved in the research knows who is eligible for the survey and who is not.

2. Find or put together a list of the target population, the list from which the sample will eventually be drawn.

3. Select the sample. Sampling methods range from simple to extremely complex. For many surveys of small populations and small areas, uncomplicated designs like simple random sampling and systematic sampling are adequate.

In the case that a population is infinite (or at least impossible to list) selecting, for example, every tenth or hundredth individual for sampling might be sufficient.

The importance of good sampling technique cannot be over emphasized. If a sampling technique makes it more likely that some members of a population will be selected than others, this can seriously skew the resulting sample's characteristics and the implications of the sample.

The sampling distribution of the sample mean

Any time you select a sample, you are likely to get different members of a population. The resulting sample mean ($\bar{x}$) will vary from sample to sample and, thus, is a random variable. As such, this random variable has a distribution we can discuss.

The sample mean is a random variable with an expected value equal to the population mean (μ) and a standard deviation equal to the population standard deviation divided by the square root of the sample size.

Put another way:

$$\mu_{\bar{x}} = E(\bar{x}) = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where

μ is the population mean

σ is the population standard deviation

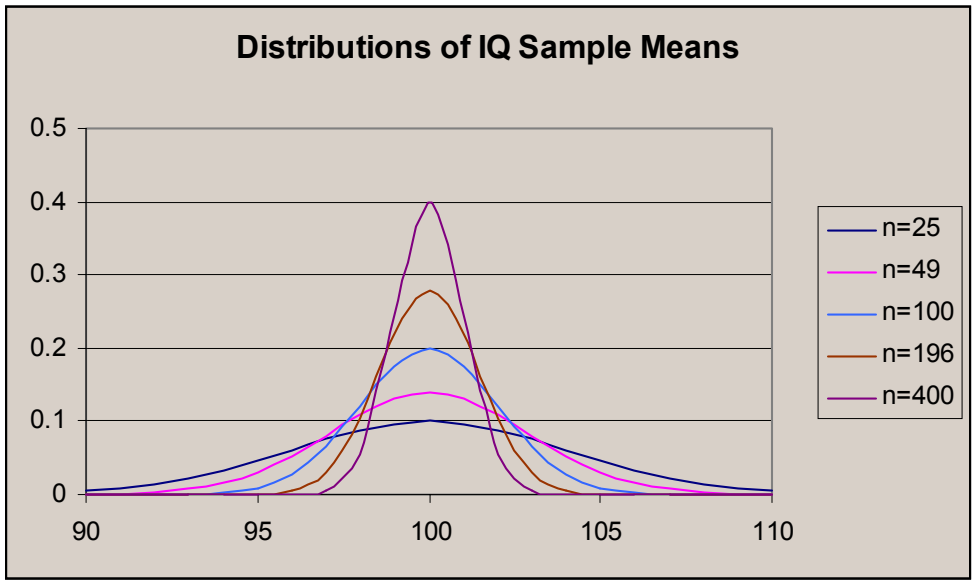
n is the sample size

EX: I.Q. tests generally have a mean of 100 and a standard deviation of 20. If such a test is administered to a population and you take a random sample of that population, what will the distribution of the sample mean score be if the sample size is 25, 49, 100, 196?

In each case, $\mu_{\bar{x}} = E(\bar{x}) = \mu = 100$. The difference is in the standard deviation of the sample mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Because the population standard deviation $\sigma = 20$, the standard deviation of the sample mean will be:

n	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
25	4.000
49	2.857
100	2.000
196	1.428
400	1.000

In a graph, these look like:



Another way to describe the distribution of the sample means is to list the probabilities that a sample mean is between, say, 98 and 102, for each sample size. To calculate these, convert the limits (98 and 102) to standard normal values by subtracting the population mean (100) and dividing by the standard deviation of the sample mean:

n	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$P(98 < \bar{x} < 102)$
25	4.000	0.3830
49	2.857	0.5160
100	2.000	0.6826
196	1.428	0.8384
400	1.000	0.9544

Central Limit Theorem
The central limit theorem says that for a sample drawn from any population (even populations with really weird distributions), if you draw enough observations for a sample then the mean of that sample will approximately follow a normal distribution. In practice, this usually means drawing at least 30 observations.

So, if you take a sample from a population, the mean of that sample will be a normally distributed random variable with expected value equal to the mean of the population and a standard deviation equal to the standard deviation of the population divided by the square root of the sample size.

EX:
* Population mean = $\mu = 200$
* Population S.D. = $\sigma = 50$
* Sample size = n = 100

The standard deviation of x-bar is
* $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5$

So, the sample mean, $\bar{x}$, is distributed normally with mean 200 and standard deviation 5.

That is, $\bar{x} \sim N(200,5)$.

What is the probability that the sample mean will be within +/- 5 of the population mean?

Find $P(195 < \bar{x} < 205)$
Convert these to standard normal random variable terms...
* (195-200)/5 = -1
* (205-200)/5 = +1

So, this is the same as...
Find $P(-1 < z < 1) = 2 \times 0.3413 = 0.6816$.

The sampling distribution of the population proportion

If the variable of interest is the percentage of the population satisfying some requirement (that they are left handed, voted in the last election or drive Cadillacs, for example) the textbook calls this population proportion *p*. The proportion of the sample satisfying the characteristic is $\bar{p}$. It's not that different from $\bar{x}$.

The sample proportion $\bar{p}$ is a random variable with expected value *p* (the population proportion) and standard deviation

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \qquad q = 1 - p$$

EX: Imagine that among the world population, 10% of the people are left handed. Find the distribution of the sample proportion which is left handed in a sample of size 81. What is the probability that, in a random sample of size 81, more than 12% of the people are left handed? That is

Find $P(\bar{p} > 0.12)$

Because the sample size 81 is fairly large, we will assume that the sample proportion will be normally distributed with mean $\mu_{\bar{p}} = 0.10$ and standard deviation

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.90 \cdot 0.10}{81}} = \sqrt{\frac{0.09}{81}} = 0.0333$$

So, we can convert the question to one involving a standard normal variable:

$$P(\bar{p} > 0.12) = P\left(z > \frac{0.12 - \mu_{\bar{p}}}{\sigma_{\bar{p}}}\right) = P\left(z > \frac{0.12 - 0.10}{0.0333}\right) = P(z > 0.6) = 0.2743$$

EX: Consider the example of a presidential election in the U.S. These elections are very expensive methods of determining the proportion of the population that wants to vote for each candidate. We could probably get the same winner most of the time by simply selecting a smaller random sample of the population and just asking them. How well would this work? Let's say that 51% of the population want the democratic candidate to win. Let p be the proportion of the population that want the democratic candidate and p=0.51.

If we take a sample of size 30, what is the probability that $\bar{p}$, the portion of the sample favoring the democratic candidate, will be greater than or equal to 0.50 (allowing the democrats to win)?

So, we're trying to find $P(\bar{p} > 0.50)$. This is a normal random variable, and to find the probability, we need to convert it to a standard normal random variable. To do this, we need the mean ($\mu_{\bar{p}} = 0.51$) and the standard deviation

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.51 \cdot 0.49}{30}} = 0.09127$$

Converting 0.50 to a comparable standard normal random variable limit yields

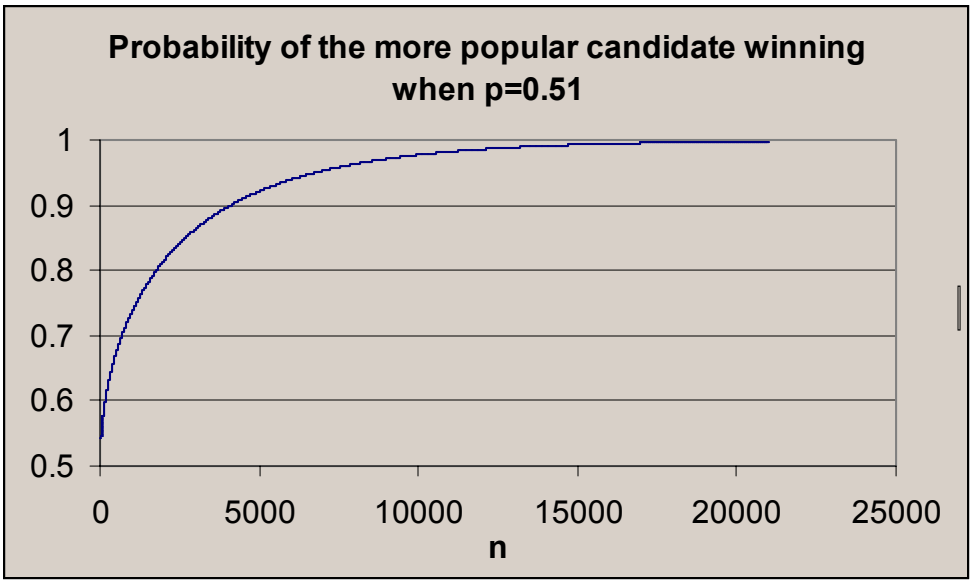
$$P(\bar{p} > 0.50) = P\left(z > \frac{0.50 - 0.51}{0.09127}\right) = P(z > -0.109) \cong 0.5438$$

So, a random sample of 30 voters would yield the correct winner about 54% of the time if 51% of the population preferred one candidate.

A random sample of 100 voters will give the correct winner 0.5793 or about 58% of the time.
A random sample of 10,000 voters will give the correct winner 0.9772 or almost 98% of the time, even when the support for the most popular candidate is at only 51%.

As a result, we can conclude that if we just asked a random sample of 10,000 people who they wanted for president, we would get the correct winner (the one with 51% support) almost 98% of the time. This would eliminate much of the expense of national elections without significantly affecting the results.

Here's graph showing the probability of selecting the correct candidate when the population proportion supporting her (or him) is 51% versus 49% for the other:



EX: A survey showed that a family spends an average of $215.60 per day while on vacation with a standard deviation of $85.00. For the following questions, assume a sample size of 40 families.
μ = 215.60 and σ = 85.00 and n=40.

A. Find the sampling distribution of the sample mean, $\bar{x}$.
$\bar{x}$ is distributed normally with mean 215.60 and standard deviation $85.00/40^{1/2}$=13.44 .

B. What is the probability that the simple random sample of 40 families will provide a sample mean that is within $20 of the population mean?

P(195.60 < $\bar{x}$ < 235.60)

(195.60-215.60)/13.44 = -20/13.44 = -1.488
(235.60-215.60)/13.44 = 20/13.44 = 1.488
P(-1.488 < z < 1.488) = 2 x 0.4319 = 0.8638.

C. What is the probability that the simple random sample of 40 families will provide a sample mean that is within $10 of the population mean?

P(205.60 < $\bar{x}$ < 225.60)

(205.60-215.60)/13.44 = -10/13.44 = -0.74
(225.60-215.60)/13.44 = 10/13.44 = 0.74
P(-0.74 < z < 0.74) = 2 x 0.2704 = 0.5408

EX: Among adults in the U.S., 17% voted for George Bush in 1992. Assume a sample of 800 adults is taken.

p=0.17, n=800.

A. What is the sampling distribution of the sample proportion of people that voted for George Bush?

$\bar{p}$ is normally distributed with mean 0.17 and standard deviation

(0.17*0.83/800)1/2 = 0.01328

B. What is the probability that the sample proportion is within two percentage points of the population proportion?

$P(0.15 < \bar{p} < 0.19)$

(0.15-0.17)/0.01328 = -0.02/0.01328 = -1.51
(0.19-0.17)/0.016 = 0.02/0.01328 = 1.51
P(-1.51 < z < 1.51) = 2 x 0.4345 = 0.8690

C. If the sample size is doubled to 1600, what is the probability that the sample proportion is within two percentage points of the population mean?

$\bar{p}$ is normally distributed with mean 0.17 and standard deviation

$(0.17*0.83/1600)^{1/2} = 0.00939$

$P(0.15 < \bar{p} < 0.19)$
(0.15-0.17)/0.00939 = -0.02/0.00939 = -2.13
(0.19-0.17)/0.00939 = 0.02/0.00939 = 2.13
P(-2.13 < z < 2.13) = 2 x 0.4834 = 0.9668

EX: A production run is not acceptable for shipment to customers if a sample of 100 items contains 5% or more defective items. If a production run has a population proportion defective of p=0.10, what is the probability that p-bar will be at least 0.05?

p=0.10, n=100

Find $P(\bar{p} > 0.05)$

$\bar{p}$ is distributed normally with mean 0.10 and standard deviation

$(0.10*0.90/100)^{1/2} = 0.03$
(0.05-0.10)/0.03 = -0.05/0.03 = -1.67
P(z > -1.67) = 1 - 0.0475 = 0.9525

So only about 4.75% of the samples will be acceptable and 95.25% will be rejected

EX: Studies show that 10% of the U.S. population is functionally illiterate. Four hundred people are selected at random to participate in a project. If more than 12% of the people selected are illiterate then the project will fail. Find the probability that this project will fail.

$\bar{p}$ is distributed normally with mean 0.10 and standard deviation

$$[(0.10)(0.90)/400]^{1/2} = 0.015$$

$P(\bar{p} > 0.12) = P(Z > (0.12-0.10)/0.015) = P(Z > 0.02/0.015) = P(Z > 1.33) = 0.0918.$