

Samples and Populations

MSc Module 6: Introduction to Quantitative Research Methods
Kenneth Benoit

February 24, 2010

Samples v. populations

- ▶ We have already introduced this terminology in Weeks 4–5: the idea that our data forms a sample, while our target of inference is a population
- ▶ The population is (generally) *unobservable*
- ▶ *Inference* means we seek to generalize about the population from observing the characteristics of a sample
- ▶ Probably the most commonly known use of sampling strategies involves the use of *surveys* of limited groups to infer characteristics of a much larger group
- ▶ Also standard in auditing strategies
- ▶ The typical, central concern with sampling is to achieve *representativeness*

Random sampling

- ▶ Basic idea: Every member of the population has an equal chance of being drawn into the sample
- ▶ Presumes that every member of the population can be identified before the sample is drawn
- ▶ Randomization is achieved in various ways:
 1. Generating numbers from chance physical events, such as drawing numbered balls, rolling dice, flipping a coin, etc.
 2. Selecting numbers from a random number table
 3. Using a computer to generate random numbers (preferred!)
- ▶ Many variations on this method of *simple random sampling* exist, since it is hard to achieve the “equal chance” standard in practice

Sampling error

- ▶ We may *expect* a sample statistic to equal a population parameter, but in each individual sample, we are likely to observe differences
- ▶ These differences will vary each time a sample is drawn, even if in the long run, the average difference will be zero
- ▶ These differences are referred to as **sampling error**: the difference caused by chance differences between the sample's characteristics and those in the population
- ▶ This notion (and even the terminology) should be familiar from polling results, which are almost always reported with a sampling error

Example: L&F Table 6.1

TABLE 6.1 *A Population and Three Random Samples of Final-Examination Grades*

Population			Sample A	Sample B	Sample C
70	80	93	96	40	72
86	85	90	99	86	96
56	52	67	56	56	49
40	78	57	52	67	56
89	49	48			
99	72	30			
96	94		$\bar{X} = 75.75$	$\bar{X} = 62.25$	$\bar{X} = 68.25$
$\mu = 71.55$					

```
> pop <- c(70, 86, 56, 40, 89, 99, 96, 80, 85, 52,
+         78, 49, 72, 94, 93, 90, 67, 57, 48, 30)
> (mu <- mean(pop))
[1] 71.55
> (s1 <- sample(pop, 4))
[1] 78 70 30 93
> mean(s1)
[1] 67.75
> (s2 <- sample(pop, 4))
[1] 90 30 89 56
> mean(s2)
[1] 66.25
> (s3 <- sample(pop, 4))
[1] 96 93 40 89
> mean(s3)
[1] 79.5
```

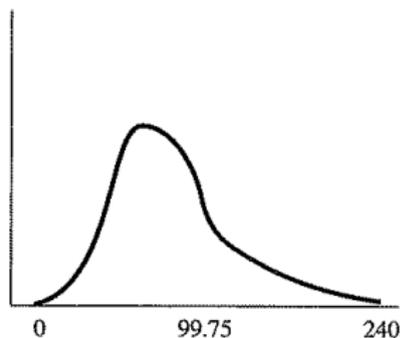
Sampling distribution of means

- ▶ Assume that we have an MSc student who can take a random telephone poll of 200 people in a day, to determine what their “probability to vote” yes is for the Lisbon Treaty referendum
- ▶ Let’s then assume that our researcher gets a large grant, and is able to repeat the survey process (with $n = 200$) a further 100 times
- ▶ The set of 100 sample means that our researcher obtained would provide a *frequency distribution*
- ▶ But using probability theory, we also know what the *probability distribution* of the sampling means will be: in particular, it will be *normally distributed*

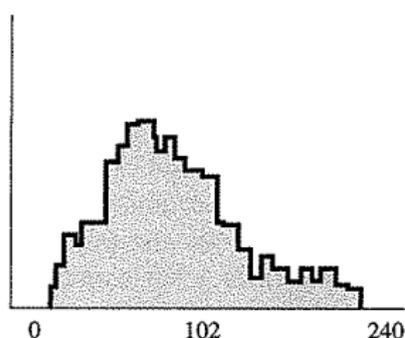
Characteristics of a sampling distribution of means

1. The sampling distribution of means will be approximately normally distributed. This is true *regardless of the distribution of the data from which the samples are drawn*.
2. The mean of a sampling distribution of means will equal the population mean.
3. The standard deviation of a sampling distribution is smaller than the standard deviation of the population. In other words, the sample mean is less variable than the scores from which it is a sample. This feature is the key to our ability to make reliable inferences from samples to populations.

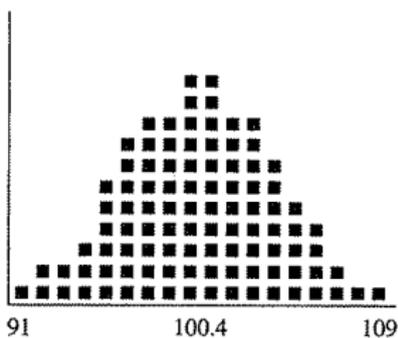
Population, sample, and sampling distributions



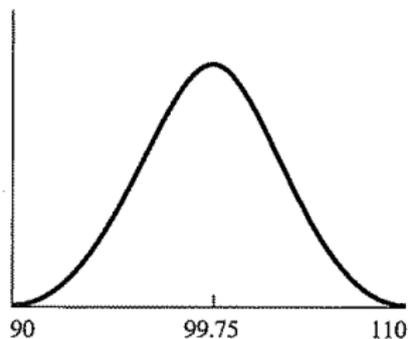
(a) Population distribution



(b) Sample distribution
(one sample with $N = 200$)



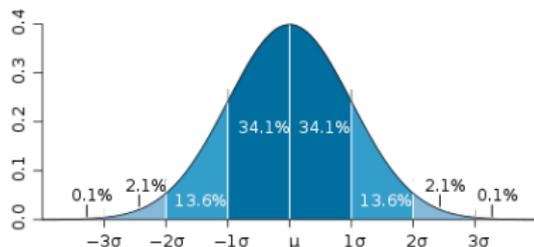
(c) Observed sampling distribution
(for 100 samples)



(d) Theoretical sampling distribution
(for infinite number of samples)

Using the normal curve to assess sample mean probabilities

- ▶ Recall that if we define probability as the likelihood of occurrence, then the normal curve can be regarded as a probability distribution
- ▶ Using the μ and σ from a normal curve, we can then assess the probability of finding specific scores along this distribution (as we did in Week 5)
- ▶ The same applies to the distribution of sampling means, since theory tells us that this will be normally distributed



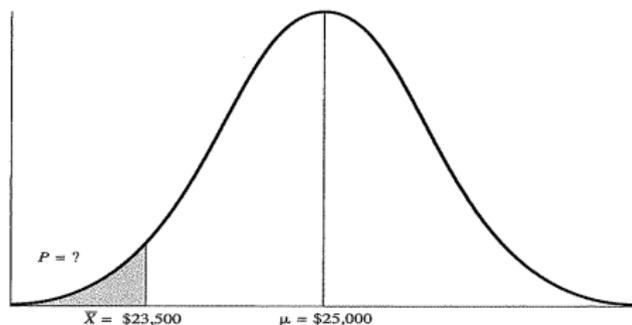
- ▶ The question: If we assume μ equals some specific value, then how likely was it to have drawn a given sample mean \bar{X} ?

Probability and the sampling distribution of means

- ▶ Remember that σ is the standard deviation of population scores
- ▶ The standard deviation of the sampling distribution (which will be smaller) is denoted as $\sigma_{\bar{x}}$
- ▶ Because the sampling distribution of means is normally distributed, we can use the z scores (Table A) to obtain the probability of obtaining any given sample mean

Sampling means example

- ▶ Imagine that UCD claims its graduates earn \$ 25,000 annually
- ▶ We decide to test this claim by sampling 100 graduates and measuring their incomes
- ▶ We obtain a sample mean of $\bar{X} = \$23,500$. How likely was it to obtain this sample mean (or less) if the true (population) earnings mean is \$ 25,000?



- ▶ Question: what is the area of the shaded region? (since this tells us the probability of obtaining a sample mean of \$ 25,000 or less)

Sampling means example cont.

1. Obtain the z score for this value, using

$$z_i = \frac{X_i - \mu}{\sigma_{\bar{x}}}$$

- ▶ \bar{x} is the sample mean (\$ 23,500)
 - ▶ μ is the mean of means (the university's claim of \$ 25,000)
 - ▶ $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of means
2. Suppose we know that the standard deviation of the sampling procedure is $\sigma_{\bar{x}} = \$700$. Then we translate to a z score as:

$$z = \frac{23,500 - 25,000}{700} = -2.14$$

Sampling means example cont.

3. Then we can consider the probability up to this value:

```
> pnorm(-2.14)
[1] 0.01617738
> round(pnorm(-2.14), 2)
[1] 0.02
```

4. What do we conclude then about the original reported value of \$ 25,000 from the university claim?

Answer: we reject the university's claim since it was very unlikely to have obtained this sample mean if the true population mean were actually \$ 25,000

Standard error of the mean

- ▶ In practice, we usually know very little about the sampling distribution of the mean, since we usually collect only a single sample
- ▶ The result is that we generally will not know $\sigma_{\bar{x}}$
- ▶ but we can derive the standard deviation of a theoretical sampling distribution (if the means of infinite samples were obtained)
- ▶ This quantity is called the **standard error of the mean**:

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

- ▶ Example: IQ test is standardized to $\mu = 100$ and $\sigma = 15$. If we have $N = 10$, then $\sigma_{\bar{x}} = 15/\sqrt{10} = 15/3.1623 = 4.74$

Confidence intervals

- ▶ Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a **confidence interval**
- ▶ Put another way, we can estimate the probability that our population mean actually falls within a range of values
- ▶ If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter
- ▶ Confidence intervals are usually calculated so that this percentage is 95%, but can be others (e.g. 99%)
- ▶ Confidence intervals are more informative than the simple results of hypothesis tests (where we decide "reject the null" or "don't reject the null"), since they provide a range of plausible values for the unknown parameter

Example

- ▶ Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46
- ▶ **Question: Is the average MEP really left of center?**
- ▶ How to approach the problem:
 - ▶ Let's use $\bar{x} = 46$ as our estimate of μ
 - ▶ Likewise, we measure $s_x = 12$, so estimate $\sigma_{\bar{x}} = 12/\sqrt{25} = 2.4$
 - ▶ We know that 95% of all values are within \pm (approximately) two standard deviations of the mean, in a normal distribution
 - ▶ In other words, the **95% C.I. = $\bar{X} \pm 1.96\sigma_{\bar{x}}$**
 - ▶ For this problem, $CI_{95} = 46 \pm 1.96 * (2.4) = [41.2, 50.8]$

Example cont.

- ▶ We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 44, give or take 4.8
- ▶ In this case, we could not conclude with 95% confidence that the average MEP is left of center, since the confidence interval includes the middle (50) value
- ▶ The specified probability level (95%) is known as the level of confidence
- ▶ Our guesses will be wrong $100\% - 95\% = 5\%$ of the time
- ▶ The precision of our estimate will be determined by the **margin of error**, which we obtain by multiplying our the standard error by the z score representing a desired level of confidence (e.g. $1.96 * 2.4 = 4.7$ in our earlier example
 - ▶ for 68%, $z = \pm 1.00$
 - ▶ for 95%, $z = \pm 1.96$
 - ▶ for 99%, $z = \pm 2.58$

The t distribution

- ▶ In most applications we never know σ to use in the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$
- ▶ Problem: If we simply substitute sample standard deviation s for σ then we will underestimate σ
- ▶ As a result, we compensate by using $N - 1$ instead of N as the denominator:

$$\hat{\sigma} = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{N - 1}}$$

- ▶ To get an unbiased estimate of the standard error of the mean from a sample, we use this formula instead:

$$s_{\bar{x}} = \frac{s}{\sqrt{N - 1}}$$

The t distribution cont.

- ▶ One result of this is that the sampling distribution of means is no longer perfectly normal – in other words $\frac{\bar{X} - \mu}{s_{\bar{x}}}$ does not quite follow the z distribution
- ▶ Instead, we call this the **t distribution**, which is like a normal distribution but slightly wider (or fatter at the tails)
- ▶ The ratio we use instead of the z ratio is called the **t ratio**:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

- ▶ Each t distribution's exact shape will also depend on the *degrees of freedom* $df = N - 1$. The greater the df , the closer t gets to the normal distribution

Probabilities and the t distribution

- ▶ Tables for area under the t distribution are represented by both degrees of freedom and different levels of α , which represent the level of confidence:

$$\alpha = 1 - \text{level of confidence}$$

- ▶ Confidence intervals for the t distribution are constructed using t scores for given degrees of freedom and α levels:

$$\text{confidence interval} = \bar{X} \pm ts_{\bar{x}}$$

Step-by-step illustration (LF&F pp197–199)

```
> ## Step-by-step illustration L&F pp131--133
> x <- c(1,5,2,3,4,1,2,2,4,3)
> mean(x)                                # step 1
[1] 2.7
> (s <- sd(x))                             # step 2
[1] 1.337494
> (s <- sqrt(sum(x^2)/10 - mean(x)^2)) # step 2 population version from L&F
[1] 1.268858
> n <- length(x)                          # step 3
> (sxbar <- s/sqrt(n-1))
[1] 0.4229526
> (tval <- qt(1-.05/2, 9))                 # step 4: t-critical value
[1] 2.262157
> (me <- sxbar * tval)                     # step 5: margin of error
[1] 0.9567852
> c(mean(x)-me, mean(x)+me)               # step 6: 95% confidence interval
[1] 1.743215 3.656785

> (tval99 <- qt(1-.01/2, 9))              # same computation for 99% CI
[1] 3.249836
> (me <- sxbar * tval99)                  # margin of error
[1] 1.374526
> c(mean(x)-me, mean(x)+me)               # 95% confidence interval
[1] 1.325474 4.074526
```

Step-by-step illustration (LF&F pp197–199) cont.

```
> t.test(x) # easier way to get t-distribution 95% CI
```

```
One Sample t-test
```

```
data: x
```

```
t = 6.3837, df = 9, p-value = 0.0001277
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
1.743215 3.656785
```

```
sample estimates:
```

```
mean of x
```

```
2.7
```

```
> t.test(x, conf.level=.99) # 99% CI
```

```
One Sample t-test
```

```
data: x
```

```
t = 6.3837, df = 9, p-value = 0.0001277
```

```
alternative hypothesis: true mean is not equal to 0
```

```
99 percent confidence interval:
```

```
1.325474 4.074526
```

```
sample estimates:
```

```
mean of x
```

```
2.7
```

Confidence intervals for proportions

- ▶ Often we may seek to estimate a population *proportion* on the basis of a random sample: e.g. proportion to vote Yes on the Lisbon Treaty referendum
- ▶ We use the a version of the standard error known as the **standard error of the proportion**:

$$s_p = \sqrt{\frac{p(1-p)}{N}}$$

where

- ▶ s_p is the standard error of the proportion
- ▶ p is the sample proportion
- ▶ N is the total number of cases in the sample
- ▶ Because proportions have only a single parameter π of which p is an estimate, we can use the normal distribution and not t
- ▶ So: 95% CI for proportions = $p \pm 1.96s_p$