API-209
Teddy Svoronos

TF Session 2
September 18, 2015

Sampling Distributions & Estimators

I. Estimators

The Importance of Sampling Randomly

Three Properties of Estimators

1. **Unbiased**

2. **Consistent**

3. **Efficient**

I am grateful to previous TFs for many of these materials.

From the Estimators Module Quiz:

Suppose you are interested in estimating the mean household income of a population and collect data on a random sample of households. Consider the following estimator:

- Estimator $= X_1$, the value of the first income listed in the sample (i.e. the income for the household we happen to have surveyed first)
- More formally, if sample incomes are $\{x_1, x_2, x_3, \ldots, x_n\}$, the estimate is x_1

If you are using this estimator to estimate the population mean, is this estimator:

- Unbiased?
- Consistent?

Practice: For all of the following questions, assume the population parameter is μ.

1. Show that $\bar{X}$ is unbiased.

2. Derive the variance of $\bar{X}$.

3. Fill out the following table:

Estimator	Unbiased?	Consistent?	Most Efficient?
$\frac{Y_1 + Y_{25} + Y_{99}}{3}$			
$\frac{\sum_{i=1}^{n}(Y_i) + 5}{n}$			
$\frac{\sum_{i=1}^{n}(Y_i)}{n}$			

Target Parameter (θ)	**Sample Size** (s)	**Point Estimator** $(\hat{\theta})$	$E(\hat{\theta})$	**Standard Deviation of Sampling Distribution** $(\sigma_{\hat{\theta}})$
μ	n	$\bar{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p}$	p	$\sqrt{\frac{p(1-p)}{n}}$
$\mu_1 - \mu_2$	n_1, n_2	$\bar{Y}_1 - \bar{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1, n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Notes:

1. The expected values and standard errors shown in the table are valid regardless of the form of the population probability density function.
2. All four estimators possess probability distributions that are **approximately normal** for **large** samples.
3. For the last two rows, the two samples are assumed to be **independent**.
4. For all rows, the samples are assumed to be **simple random samples**.

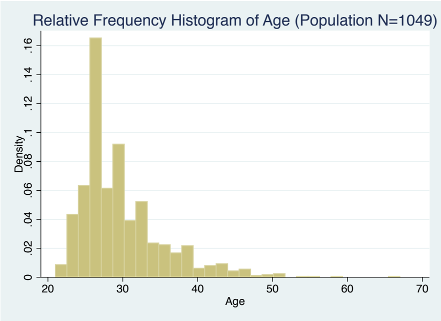
II. The Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ constitute a simple random sample from a population with mean μ and variance σ^2, then the sample mean $\overline{X}$ has an *approximately normal distribution* with mean μ and standard error $\sigma / \sqrt{n}$, assuming the sample size is large enough (usually $n > 30$ is sufficient). The approximation improves as n increases.

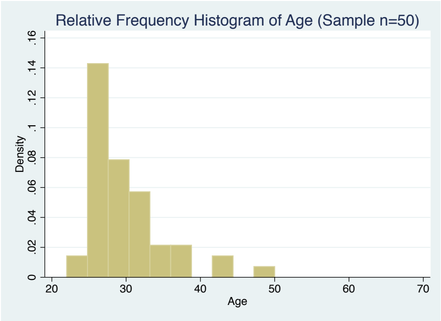
This result holds regardless of the form of the population distribution.

See this applet to visualize the CLT: http://onlinestatbook.com/stat_sim/sampling_dist/
Experiment with different sample sizes and population distributions to see how the CLT is invoked. Keep in mind this distinction from lecture:

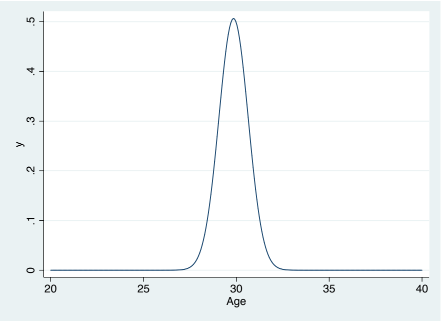
Distribution in the Population



Distribution in the Sample



"Sampling" Distribution

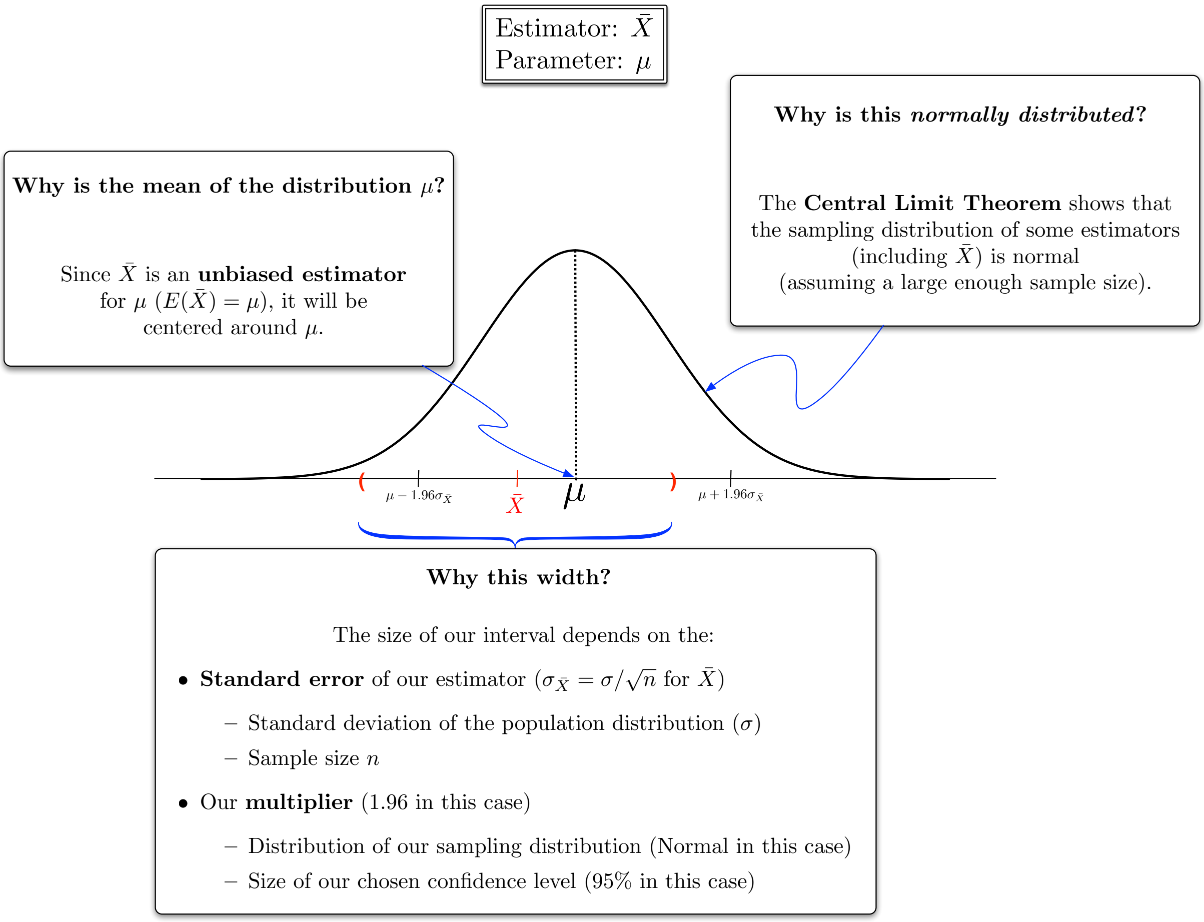


Example: Survey of Registered Voters

A telephone survey of 2374 adult Americans finds that 1912 of the respondents are registered voters. Construct a 90% Confidence Interval for the proportion of adults in the United States who are registered voters.

III. Bringing it All Together

Means



Estimator: $\bar{X}$
Parameter: μ

Why is the mean of the distribution μ?

Since $\bar{X}$ is an **unbiased estimator** for μ ($E(\bar{X}) = \mu$), it will be centered around μ.

Why is this *normally distributed*?

The **Central Limit Theorem** shows that the sampling distribution of some estimators (including $\bar{X}$) is normal (assuming a large enough sample size).

Why this width?

The size of our interval depends on the:

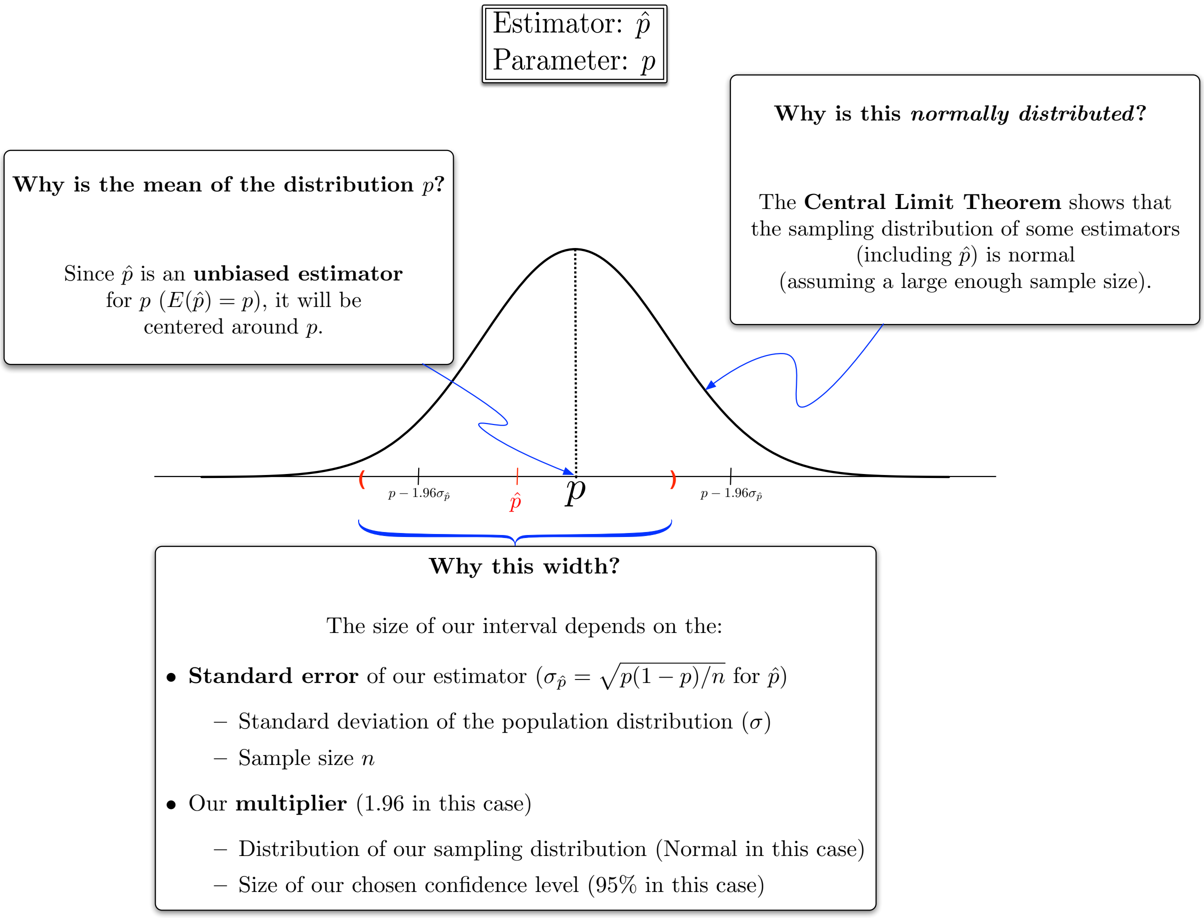
- **Standard error** of our estimator ($\sigma_{\bar{X}} = \sigma/\sqrt{n}$ for $\bar{X}$)
  - Standard deviation of the population distribution (σ)
  - Sample size n
- Our **multiplier** (1.96 in this case)
  - Distribution of our sampling distribution (Normal in this case)
  - Size of our chosen confidence level (95% in this case)

Why this particular estimate of $\bar{X}$? **No reason!**

It's just the estimate that we happened to get from a single sample of the population.

In real life, this value (and its standard deviation) is all we have to work with.

Proportions



Estimator: $\hat{p}$
Parameter: p

Why is the mean of the distribution p?

Since $\hat{p}$ is an **unbiased estimator** for p ($E(\hat{p}) = p$), it will be centered around p.

Why is this *normally distributed*?

The **Central Limit Theorem** shows that the sampling distribution of some estimators (including $\hat{p}$) is normal (assuming a large enough sample size).

Why this width?

The size of our interval depends on the:

- **Standard error** of our estimator ($\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ for $\hat{p}$)
  - Standard deviation of the population distribution (σ)
  - Sample size n
- Our **multiplier** (1.96 in this case)
  - Distribution of our sampling distribution (Normal in this case)
  - Size of our chosen confidence level (95% in this case)

Why this particular estimate of $\hat{p}$? **No reason!**

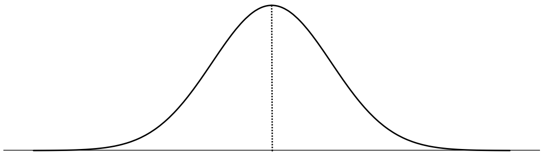
It's just the estimate that we happened to get from a single sample of the population.

In real life, this value (and its standard deviation) is all we have to work with.

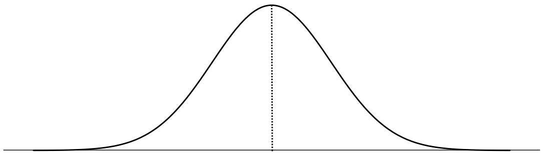
IV. Confidence Intervals

Think about the various factors that contribute to a confidence interval of the mean age of a population. What would happen to the sampling distribution (e.g. would it change shape? center?) and confidence interval (e.g. would it get larger? smaller?) as the following **increased**?

1. Sample size



2. Standard deviation of age in the population



3. Confidence level