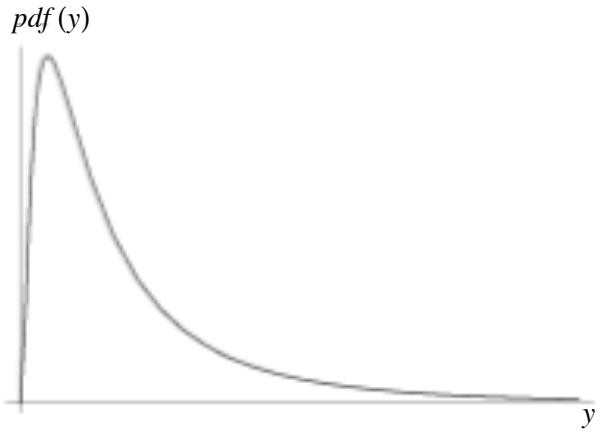
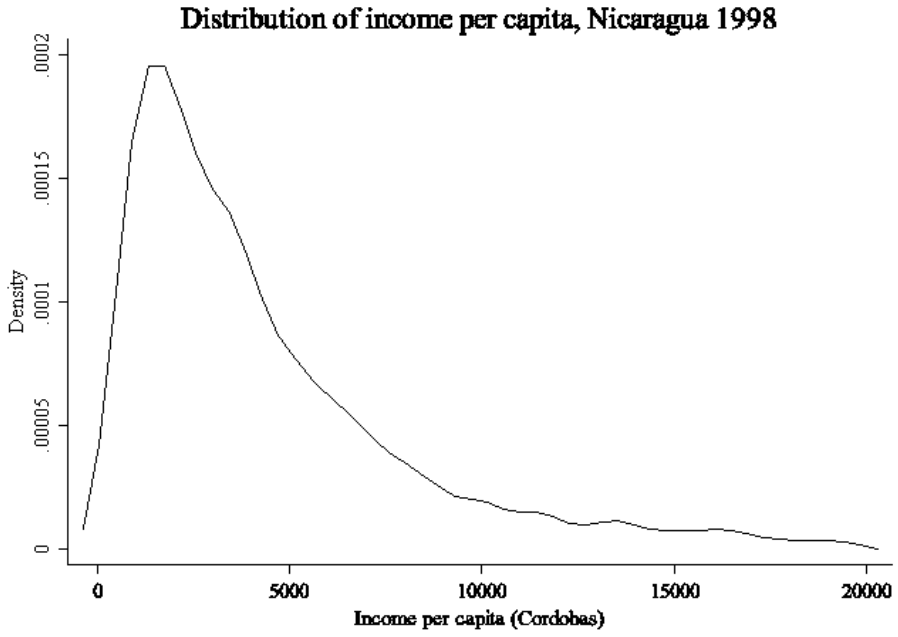
Estimation of the mean of a population and construction of a confidence interval

Consider the population of all households in Nicaragua. The distribution of income per capita x in this population is characterized by its (unknown) mean, $E(y)$ or μ, and variance, var(y) or σ^2. Its probability density function probably looks like this:



I am using a very large survey of 23,000 households in 1998 to show an approximation of the population density:

```
label var i_income "Income per capita (Cordobas)"
kdensity i_income if i_income<20000, title ("Distribution of income per
capita, Nicaragua 1998")
```



Can we use a sample to obtain an estimation of this mean μ

1. Define an estimator for the population mean μ : $\bar{y}$

One sample of 2,234 persons:

Compute sample mean and sample variance of income/capita

```
. sum income
  Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    income |      2234        5396.51   8011.15           0   168552.9
```

$$\bar{y} = \frac{\sum y_i}{n} = 5397 \qquad s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = 8011^2$$

2. Property of the estimator $\bar{y}$

It is unbiased, i.e., $E(\bar{y}) = \mu$,

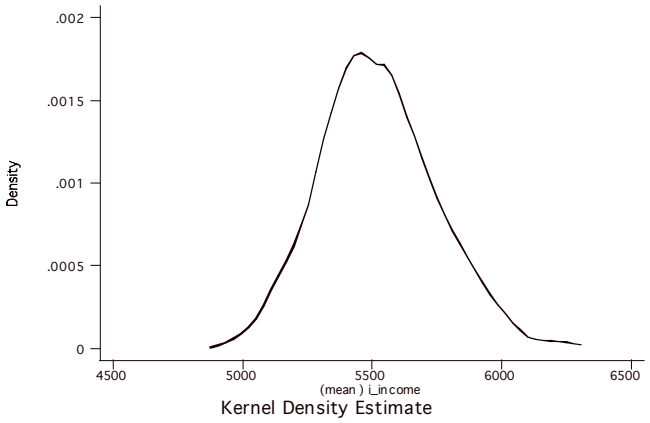
its variance is $\text{var}(\bar{y}) = \frac{1}{n}\text{var}(y) = \frac{1}{n}\sigma^2$ or standard deviation: $\text{sd}(\bar{y}) = \frac{1}{\sqrt{n}}\text{sd}(y)$

If n is large, it follows a Normal distribution $\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Check numerically:

Repeat 512 times: Here are some values for $\bar{y}$:

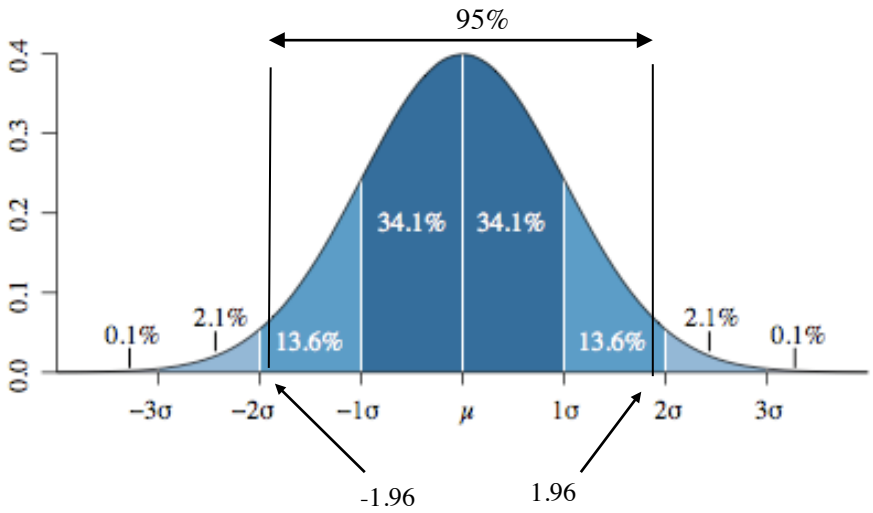
5546.19	5524.84	5608.63	5656.64	5009.88	5723.81	5281.5
5471.63	5598.41	5349.31	5462.35	5515.2	5677.64	5995.86
5503.55	5160.92	5976.73	5491.89	5641.75	5484.19	5460.17
5703.7	5384.34	5384.08	5976.77	5611.49	5272.66	5178.83
5280.9	5324.98	5394.74	5759.59	5558.9	5622.06	5548.91
5707.15	5679.66	5133.75	5717.11	5584.75	5509.09	5559.16
5157.59	5846.83	5118.19	5820.1	5653.99	5130.67	5377.95
5686.97	5493.53	5810.17	5265.17	5495.76	5442.15	5665.51
5490.38	5498.91	5997.88	5150.2	5578.23	5426.48	5669.8
5061.16	5862.29	5446.75	5484.8	5971.43	5173.89	5364.94
5396.54	5963.28	5508.36	5901.36	5383.45	5299.95	5555.01
5851.09	5914.9	5308.24	5759.63	5571.92	5160.21	5602.39
5749.47	5387.94	5474.79	5524.78	5553.74	5377.61	5295.98
5514.25	5402.17	5501.91	5646	5377.47	5533.67	5512.72
5433.35	5507.5	5398.02	5117.08	5917.58	5493.15	5497.06



It looks like a Normal distribution.

3. Reminder on the Normal distribution:

if $x \sim N(\mu,\sigma^2)$, then $\frac{x-\mu}{\sigma} \sim N(0,1)$



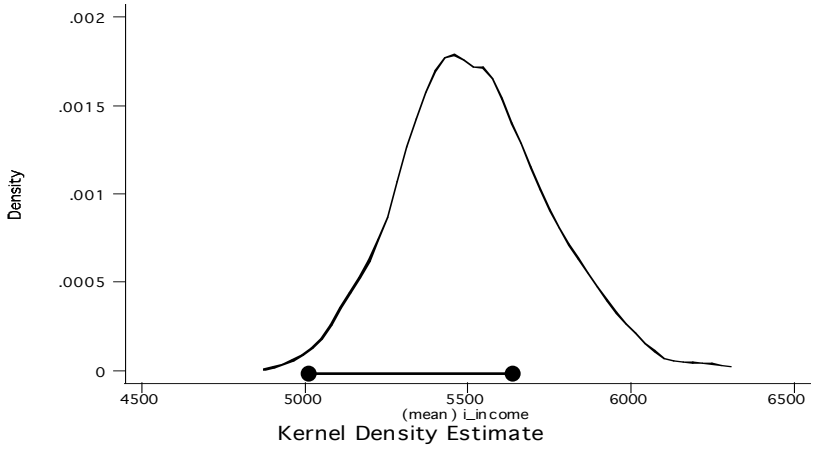
4. Compute a confidence interval for the true value μ

If you don't know the true value of σ, then use *s*, and the Student *t* distribution.
When *n* very large, *t* distribution is the same as Normal

Then, construct intervals that have 95% chance to contain μ

	$\bar{y}$	s	$\bar{y}-1.96*s/\sqrt{n}$	$\bar{y}+1.96*s/\sqrt{n}$
1	5397	8011	5064	5729
2	5283	8817	4918	5649
3	5434	7596	5119	5749
4	5403	7406	5096	5710
5	5510	14259	4919	6101
6	5697	9979	5283	6111
7	5331	8408	4983	5680
8	6027	14905	5409	6645
9	5386	9455	4993	5778
10	5123	7188	4825	5421
11	5688	14232	5097	6278
12	5597	15076	4972	6223
13	5440	9030	5066	5814
14	5471	9091	5094	5848
15	5537	14116	4952	6123

What does the confidence interval mean?



5. Special case of a binary variable

x only takes values 0 or 1.
Let *p* be the *true but unknown* proportion of 1 in the population.

For one random observation: $E(x) = p$ and $\text{var}(x) = p(1-p)$

If one collects a sample of *n* observations, and compute the average $\bar{x}$ of these observations, then

$$E(\bar{x}) = p \quad \text{and} \quad \text{var}(\bar{x}) = \frac{p(1-p)}{n}$$

Example (CNN-ORC poll)
Based on 305 registered voters who describe themselves as republicans and 139 who describe themselves as independents who lean republican, for a total of 444 republicans -- sampling error: +/- 4.5 percentage pts.

I'm going to read a list of people who may be running in the republican primaries for president in 2016. After I read all the names, please tell me which of those candidates you would be most likely to support for the republican nomination for president in 2016, or if you would support someone else. Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Jim Gilmore, Lindsey Graham, Mike Huckabee, Bobby Jindal, John Kasich, George Pataki, Rand Paul, Marco Rubio, Rick Santorum, Donald Trump, Or Scott Walker. (random order)

Sept 17-19 2015 : Trump 24%

Let's say the question was: Would you vote for Donald Trump? Yes/No
The true proportion in the population of "*republican or leaning republican registered voters"* would vote is *p*

Estimate for *p*: $\bar{x} = .24$
Estimate of the variance of $\bar{x}$: $\widehat{var(\bar{x})} = \frac{.24*.76}{444} = .00041$
$se(\bar{x}) = \sqrt{.00041} = .02$
"Sampling error" is 4 percentage points

95% confidence interval for *p*: $[\bar{x} - 1.96 * se(\bar{x}), \bar{x} + 1.96 * se(\bar{x})] = [.201, .279]$