Sampling Distribution of $\bar{X}$ and Simulation Methods

Lecture 11

Reading: Sections 10.3 – 10.5

1

Ontario Public Sector Salaries

- Public Sector Salary Disclosure Act, 1996
  - Requires organizations that receive public funding from the Province of Ontario to disclose annually the names, positions, salaries and total taxable benefits of employees paid \$100,000 or more in a calendar year
    - E.g. Government of Ontario, Crown Agencies, Municipalities, Hospitals, Boards of Public Health, School Boards, Universities, Colleges, Hydro One, Ontario Power Generation, etc.

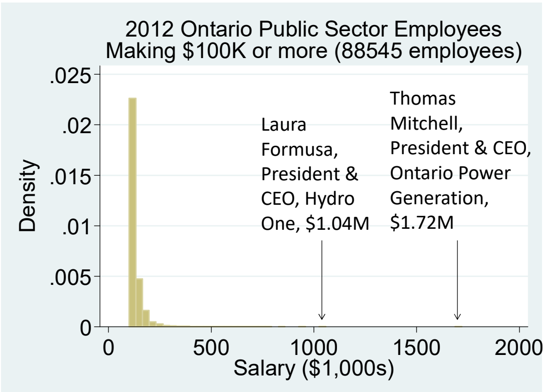
2013 disclosure of 2012 salaries: https://www.ontario.ca/page/public-sector-salary-disclosure-act-disclosures-2013

2

Sampling Error a Plausible Explanation for $\bar{X}$ being \$3,700 above μ?

- For *all* ON public sector employees w/ salaries of \$100K+, mean is \$127.5K and s.d. \$39.6K
  - Are these numbers parameters or statistics?
  - Shape of the salary distribution? (2 explanations)
- Random sample of 1,000 Ontario public sector employees has a mean salary of \$131.2K
  - Why is $\bar{X}$ different than μ?
    - How likely is *such a big* sample mean if claim true? i.e. $P(\bar{X} \geq 131.2 \mid \mu = 127.5, \sigma = 39.6, n = 1{,}000) = ?$

3



4

STATA Summary of Population

```
                          salary
-------------------------------------------------------------
      Percentiles      Smallest
 1%      100.168            100
 5%     100.9921            100
10%     102.0471            100       Obs               88545
25%     105.7447            100       Sum of Wgt.       88545

50%     115.3013                      Mean           127.5176
                        Largest       Std. Dev.      39.64454
75%     133.2821        843.095
90%     164.5416       935.2365       Variance        1571.69
95%      193.125        1036.74       Skewness       5.019101
99%     296.8753           1720       Kurtosis       64.99817
```

Note: Technically, $\sigma = 39.6443$. STATA computes s, not σ: but degrees of freedom correction matters little given large number of observations.

5

Mean and Variance of $\bar{X}$

- $\mu_{\bar{X}} = E[\bar{X}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$
- $\sigma_{\bar{X}}^2 = V[\bar{X}] = V\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

In deriving $\sigma_{\bar{X}}$ above, why is $V[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} V[X_i]$?

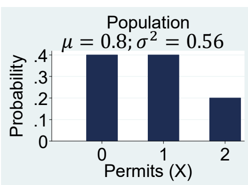
6

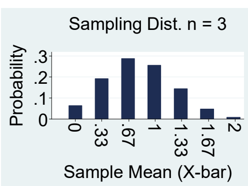
10% Condition / 10% Rule

- Derivation of $\sigma_{\bar{X}}^2$ assumes that each observation (X_i) is *independent* of others
  - For this to be true, must sample *with replacement* OR sample *without replacement* from a population that is infinitely large
  - In contrast, real applications involve sampling without replacement from a finite population
  - BUT if sample < 10% of population, independence assumption is true enough: can use $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

7

Recall Parking Permit Ex (Lec. 10)





$E[\bar{X}] = \mu = 0.8 = 0 * 0.064 + \frac{1}{3} * 0.192 + \frac{2}{3} * 0.288 + 1 * 0.256 + \frac{4}{3} * 0.144 + \frac{5}{3} * 0.048 + 2 * 0.008$

$V[\bar{X}] = \frac{\sigma^2}{n} = \frac{0.56}{3} = 0.187$
$= (0 - 0.8)^2 0.064 + \left(\frac{1}{3} - 0.8\right)^2 0.192 + \left(\frac{2}{3} - 0.8\right)^2 0.288 + (1 - 0.8)^2 0.256 + \left(\frac{4}{3} - 0.8\right)^2 0.144 + \left(\frac{5}{3} - 0.8\right)^2 0.048 + (2 - 0.8)^2 0.008$

Work to find $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}^2$ not needed. Why is work needed?

8

Shape of sampling distribution of $\bar{X}$?

- $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, but shape: $\bar{X} \sim$?
  - Central Limit Theorem (CLT): For a random sample from *any* population the sampling distribution of the sample mean ($\bar{X}$) is approximately Normal for a sufficiently large sample size
    - *Rough* rule of thumb: $n \geq 30$. But, $n < 30$ sufficient for mildly non-Normal populations: $n = 1$ is sufficient for Normal populations. Further, $n > 60$ (or more) may be required for very skewed populations.

9

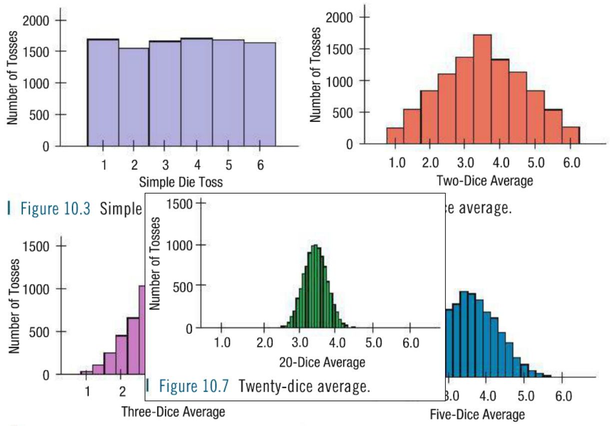
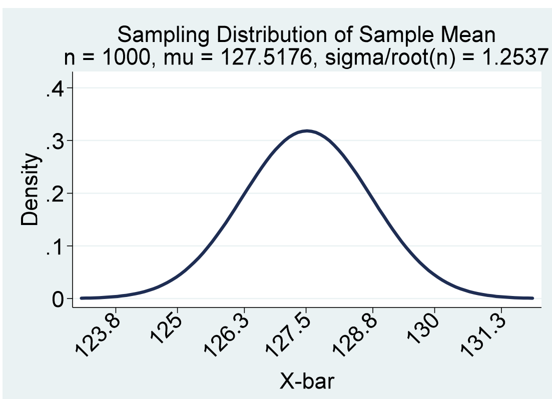


Figure 10.5 Three-dice average. | Figure 10.6 Five-dice average.



Is sampling error a *plausible* explanation for $\bar{X}$ as big as 131.2?

Sampling Error: Plausible Explanation?

$$P(\bar{X} \geq 131.2 \mid \mu = 127.518, \sigma = 39.645, n = 1{,}000)$$

$$= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{131.2 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$$

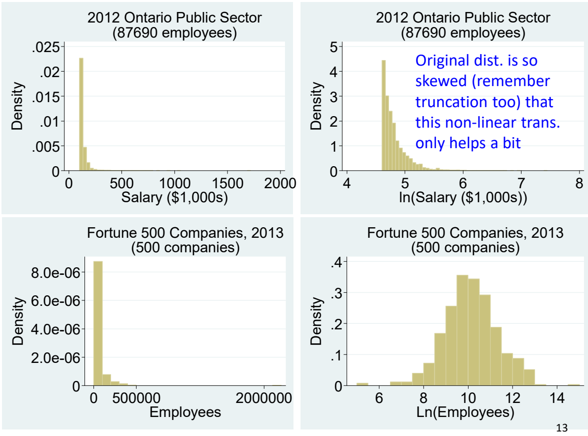
$$= P\left(Z \geq \frac{131.2 - 127.518}{39.645/\sqrt{1{,}000}}\right)$$

$$= P\left(Z \geq \frac{3.682}{1.254}\right)$$

$$= P(Z \geq 2.94) = 0.0016$$

What if sample size 50?
$P(\bar{X} \geq 131.2 \mid \mu = 127.5,$
$\sigma = 39.6, n = 50) = ?$

Which *serious* issue may we face in trying to find this probability?



13

Monte Carlo Simulation

- Monte Carlo Simulation: A problem solving method where a computer generates many random samples and you make an inference based on patterns in outcomes
  - Simulation is most useful when theoretical results (e.g. CLT) do not apply and the problem is too big for an analytic approach
- It allows us to find sampling distributions with a high degree of accuracy

14

Recall Central Limit Theorem

- The CLT says the sampling distribution of the sample mean is Bell shaped no matter what the shape of the population so long as the sample size is sufficiently large
  - What is sufficiently large?
  - Is a "rule of thumb" always correct or is it just a rough guide?
  - What factors affect how large is sufficiently large?

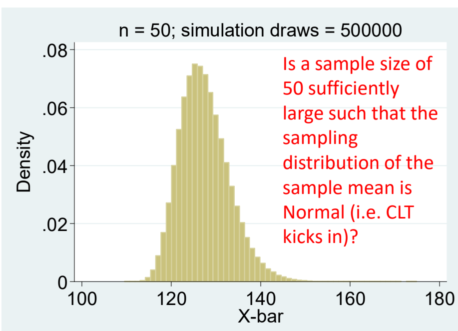
15

n = 50: Sufficiently Large?

- Monte Carlo simulation: many samples of 50 ON public employees (in each sample, n = 50)
  - # simulation draws (# samples drawn) = very big
  - Simulation error: Chance difference between simulated probability and true probability
    - Drive it to zero by doing many draws
  - For each sample compute the sample mean
  - Summarize distribution of $\bar{X}$: graphically (histogram) and numerically (Stata summary)

16

Simulated Sampling Distribution of $\bar{X}$ for $n = 50$



17

Simulated Sampling Dist. of $\bar{X}$, $n = 50$

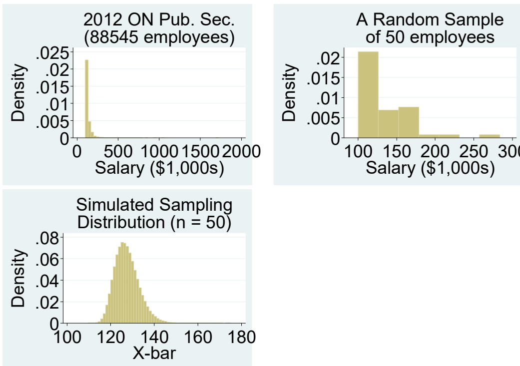
```
                            X-bar
-------------------------------------------------------------
      Percentiles      Smallest
 1%     116.9729       109.5587
 5%     119.4248       109.6845
10%     120.8754       111.0465       Obs              500000
25%     123.5441       111.2133       Sum of Wgt.      500000

50%     126.9508                      Mean            127.513
                        Largest       Std. Dev.      5.600294
75%     130.8465       172.6622
90%     134.8423       173.6038       Variance        31.3633
95%     137.4918        174.159       Skewness       .6994546
99%     143.1469       174.9272       Kurtosis       4.167933
```

Recall that $\mu \approx \$127.5K$ and $\sigma \approx \$39.6K$ in the population. What is meaning of $\sigma_{\bar{X}} = \frac{39.6}{\sqrt{50}}$ and where does it appear above?

Is sampling error a plausible explanation for an $\bar{X}$ above \$132K?

18

Three Very Different Histograms



19

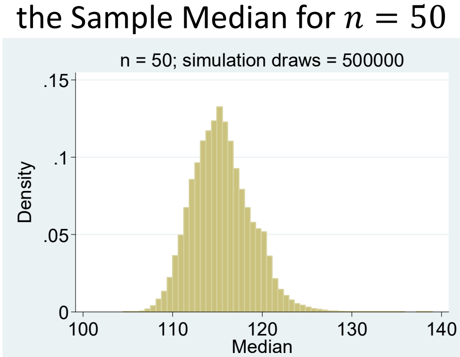
Summary of a Random Sample

salary

	Percentiles	Smallest		
1%	100.1664	100.1664		
5%	100.9522	100.9473		
10%	102.0943	100.9522	Obs	50
25%	108.7771	101.021	Sum of Wgt.	50
50%	121.4592		Mean	132.7467
		Largest	Std. Dev.	34.22585
75%	155	173.4973		
90%	167.9037	183.4379	Variance	1171.409
95%	183.4379	219.4789	Skewness	2.125154
99%	283.6693	283.6693	Kurtosis	9.144829

20

Simulated Sampling Distribution of the Sample Median for $n = 50$



21

Simulated Sampling Distribution of the Sample Median, $n = 50$

Median

	Percentiles	Smallest		
1%	108.8332	104.4422		
5%	110.5338	104.7897		
10%	111.4963	104.8258	Obs	500000
25%	113.2028	104.97	Sum of Wgt.	500000
50%	115.2876		Mean	115.4981
		Largest	Std. Dev.	3.265556
75%	117.5475	135.461		
90%	119.9086	137.6988	Variance	10.66386
95%	121.0002	138.1573	Skewness	.4225524
99%	124.086	139.0575	Kurtosis	3.392273

Recalling that the population median is \$115.3013K, is sampling error a plausible explanation for a sample median above \$118K?

How about above \$136K?

22