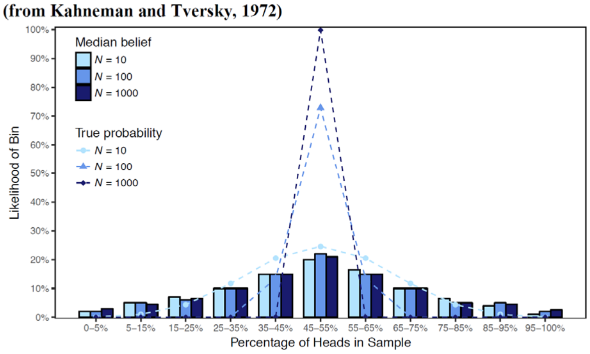
Sampling Distributions and the Sampling Distribution of $\hat{P}$

Lecture 10

Reading: Sections 10.1 – 10.2

1

Figure 1a. Sample-size neglect for binomial with rate $\theta = 0.5$ (from Kahneman and Tversky, 1972)



Source: Benjamin (2018) "Errors in probabilistic reasoning and judgement biases"
https://www.nber.org/papers/w25200.pdf

2

Sampling Distributions

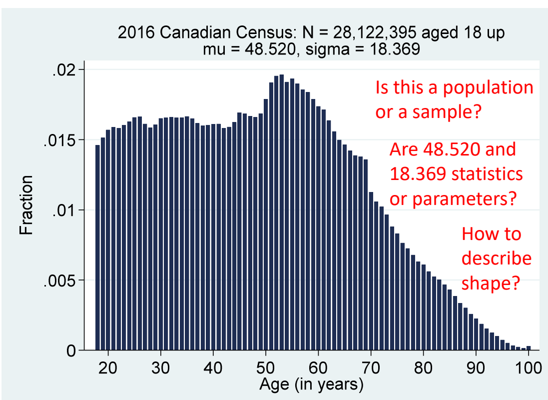
- Sampling distribution: The distribution of a sample statistic
  - Thought experiment: imagine repeated sampling
  - Sample statistics (e.g. $\bar{X}$, $\hat{P}$) are random variables
  - Distribution due to sampling error
    - Variability of the distribution measures sampling noise
  - Can be discrete or continuous: depends on population, statistic, sample size
    - Sample median vs. mean cats per household?

3

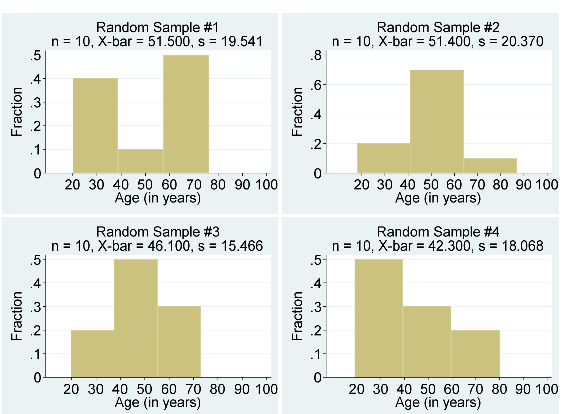
Finding Sampling Distributions

- Analytically: Figure out probability of every possible value of the sample statistic
  - Use probability rules
- With theoretical results: Central Limit Theorem; Laws of Expectation & Variance
- With simulation: Draw many samples and observe the relative frequency of each value of the sample statistic

4



Source: Statistics Canada, 2016 Census of Population, Catalogue no. 98-400-X2016008 5



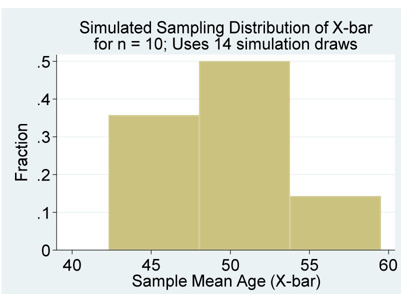
6

Mini Simulation Results: $n = 10$

	Sample Number													
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14
1	76	43	41	80	78	39	73	49	75	57	68	19	54	45
2	34	59	59	31	50	83	50	30	51	59	29	56	82	36
3	68	18	38	19	56	52	74	61	68	69	29	39	33	57
4	66	47	20	40	58	23	42	80	29	48	59	47	85	57
5	65	60	62	37	39	59	41	52	46	78	60	34	78	30
6	69	60	73	49	39	81	59	81	18	72	46	40	63	59
7	47	59	39	25	36	38	28	28	41	62	78	72	28	66
8	20	21	33	48	58	38	22	57	18	64	35	52	56	32
9	34	87	50	61	49	24	41	31	45	57	48	44	25	40
10	36	60	46	33	42	26	37	68	53	29	66	83	43	64
$\bar{X}$	51.5	51.4	46.1	42.3	50.5	46.3	46.7	53.7	44.4	59.5	51.8	48.6	54.7	48.6

7

Graph Mini Simulation Results

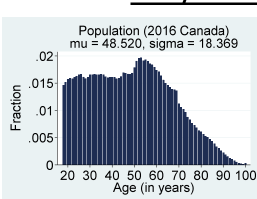


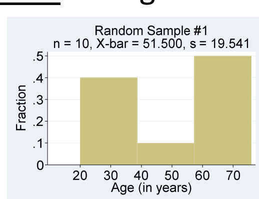
In a random sample of 10 Canadian adults (18+) in 2016, would it be surprising to get a mean age above 53?

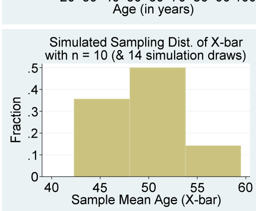
Surprising to get a mean age below 35?

8

Three <u>Very Different</u> Histograms







9

Derive the Sampling Distribution

- To analytically find the sampling distribution of a sample statistic:
  1. List every sample with n observations that is possible from the population of interest
  2. Find the probability of obtaining each possible sample
  3. Calculate the sample statistic of interest for each possible sample
  4. Link values in 3. with probabilities in 2.

10

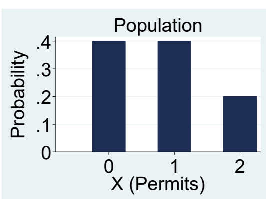
Parking Permit Example

- A town has a street parking shortage
  - 500,000 households in town
  - Town starts requiring parking permits and caps number per household to 2
    - Each gets 0, 1, or 2 permits (1 per vehicle)
    - Issues 400,000 permits: mean = 0.8 per household
- Paul does a survey, n = 3, asks # permits
  - Finds sample mean of 2: all 3 have 2 permits
  - Why a discrepancy between 2 and 0.8?

11

Town's Report on Permit Distribution

# permits	Prop. of residents
0	0.4
1	0.4
2	0.2



$\mu = E[X] = \sum xp(x)$

$\mu = 0 * 0.4 + 1 * 0.4 + 2 * 0.2 = 0.8$

$\sigma^2 = V[X] = \sum (x-\mu)^2 p(x)$

$\sigma^2 = (0-0.8)^2 0.4 + (1-0.8)^2 0.4 + (2-0.8)^2 0.2 = 0.56$

$\sigma = \sqrt{\sigma^2} = \sqrt{0.56} = 0.748$

12

1. List All Possible Samples (n=3)

sample	sample	sample
0,0,0	1,0,0	2,0,0
0,0,1	1,0,1	2,0,1
0,0,2	1,0,2	2,0,2
0,1,0	1,1,0	2,1,0
0,1,1	1,1,1	2,1,1
0,1,2	1,1,2	2,1,2
0,2,0	1,2,0	2,2,0
0,2,1	1,2,1	2,2,1
0,2,2	1,2,2	2,2,2

How many different samples are possible?

13

2. Probability of Each Sample

sample	probability	sample	probability	sample	probability
0,0,0	$(.4)^3$=.064	1,0,0	$(.4)^3$=.064	2,0,0	$(.2)(.4)^2$=.032
0,0,1	$(.4)^3$=.064	1,0,1	$(.4)^3$=.064	2,0,1	$(.2)(.4)^2$=.032
0,0,2	$(.2)(.4)^2$=.032	1,0,2	$(.2)(.4)^2$=.032	2,0,2	$(.2)^2(.4)$=.016
0,1,0	$(.4)^3$=.064	1,1,0	$(.4)^3$=.064	2,1,0	$(.2)(.4)^2$=.032
0,1,1	$(.4)^3$=.064	1,1,1	$(.4)^3$=.064	2,1,1	$(.2)(.4)^2$=.032
0,1,2	$(.2)(.4)^2$=.032	1,1,2	$(.2)(.4)^2$=.032	2,1,2	$(.2)^2(.4)$=.016
0,2,0	$(.2)(.4)^2$=.032	1,2,0	$(.2)(.4)^2$=.032	2,2,0	$(.2)^2(.4)$=.016
0,2,1	$(.2)(.4)^2$=.032	1,2,1	$(.2)(.4)^2$=.032	2,2,1	$(.2)^2(.4)$=.016
0,2,2	$(.2)^2(.4)$=.016	1,2,2	$(.2)^2(.4)$=.016	2,2,2	$(.2)^3$=.008

14

3. Mean of Each Sample

sample	mean	sample	mean	sample	mean
0,0,0	0	1,0,0	0.33	2,0,0	0.67
0,0,1	0.33	1,0,1	0.67	2,0,1	1
0,0,2	0.67	1,0,2	1	2,0,2	1.33
0,1,0	0.33	1,1,0	0.67	2,1,0	1
0,1,1	0.67	1,1,1	1	2,1,1	1.33
0,1,2	1	1,1,2	1.33	2,1,2	1.67
0,2,0	0.67	1,2,0	1	2,2,0	1.33
0,2,1	1	1,2,1	1.33	2,2,1	1.67
0,2,2	1.33	1,2,2	1.67	2,2,2	2

How many different means are possible?

15

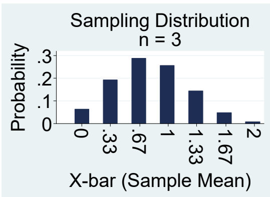
4. Probability of Each Mean

mean	probability
0	0.064
0.33	0.192 = .064 + .064 + .064
0.67	0.288 = .032 + .064 + .032 + .064 + .064 + .032
1	0.256 = .032 + .032 + .032 + .064 + .032 + .032 + .032
1.33	0.144 = .016 + .032 + .032 + .016 + .032 + .016
1.67	0.048 = .016 + .016 + .016
2	.008

16

Sampling Distribution of Mean

mean	probability
0.00	0.064
0.33	0.192
0.67	0.288
1.00	0.256
1.33	0.144
1.67	0.048
2.00	0.008



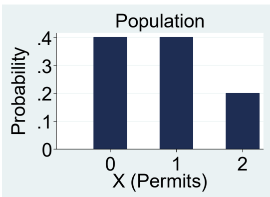
$E[\bar{X}] = 0 * 0.064 + \cdots + 2 * 0.008 = 0.8$

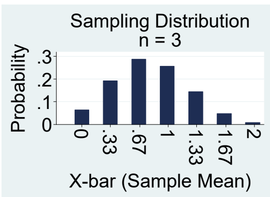
$V[\bar{X}] = (0 - 0.8)^2 0.064 + \ldots + (2 - 0.8)^2 0.008 = 0.187$

Discrete or continuous? $SD[\bar{X}] = \sqrt{0.187} = 0.432$

17

Population Dist. ≠ Sampling Dist.

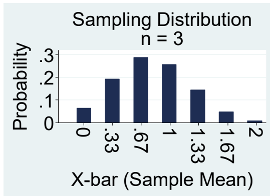




18

Explanation for Discrepancy?

- Why is $\bar{X} \neq \mu$? I.e. $\bar{X} = 2$ but $\mu = 0.8$.
- Potential explanations:
  - Sampling error
  - Non-sampling errors
  - Parameter is not what it is claimed to be
- Which explanations are *plausible* in permit example?



Scanning Cargo Example

- Bloomberg "U.S. Backs Off All-Cargo Scanning Goal With Inspections at 4%" Aug 13, 2012
  - "Customs and Border Protection officials scanned with X-ray or gamma-ray machines 473,380, or 4.1 percent, of the 11.5 million containers shipped in the fiscal year ended Sept. 30"
- Suppose a shipping company sent 10,000 containers during that period and 4.31 percent were scanned: why is 4.31 > 4.1?

https://www.bloomberg.com/news/articles/2012-08-13/u-s-backs-off-all-cargo-scanning-goal-with-inspections-at-4-

Inference about Proportions

- Population proportion (a parameter): p
- Sample proportion (a statistic): $\hat{P} = \frac{X}{n}$
- Use $\hat{P}$ to make an inference about p
  - e.g. What proportion will vote for a candidate?
- We know a lot about $\hat{P}$: X is Binomial $X \sim B(n, p)$ and $\frac{X}{n}$ is a linear transformation (i.e. doesn't change shape)

Sampling Distribution of $\hat{P}\ \left(=\frac{X}{n}\right)$

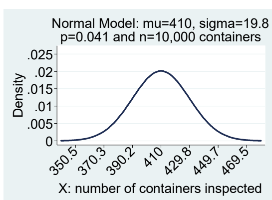
- $X\sim$ Binomial but may be approx. Normal if:
  $np \pm 3\sqrt{np(1-p)}$ is between 0 and n
  OR $np \geq 10$ and $n(1-p) \geq 10$
- $E[\hat{P}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}np = p$
- $V[\hat{P}] = V\left[\frac{X}{n}\right] = \frac{1}{n^2}V[X] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$
- $\hat{P}\sim N\left(p, \frac{p(1-p)}{n}\right)$ when n is sufficiently large

What does $SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}}$ tell us?

22

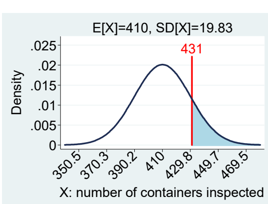
Cargo Example: X

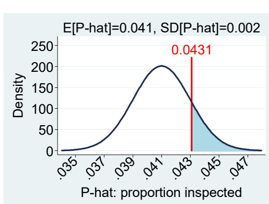
- X: # containers inspected
  - $p = 0.041, n = 10{,}000$
  - $E[X] = 10000 * 0.041 = 410$
  - $V[X] = 10000 * 0.041 * 0.959 = 393.2$
  - $SD[X] = \sqrt{393.2} = 19.8$
  - X is Binomially distributed, but is Normal approximation good?



23

Of 10,000 containers, 431 (4.31%) inspected: Why were so many inspected?

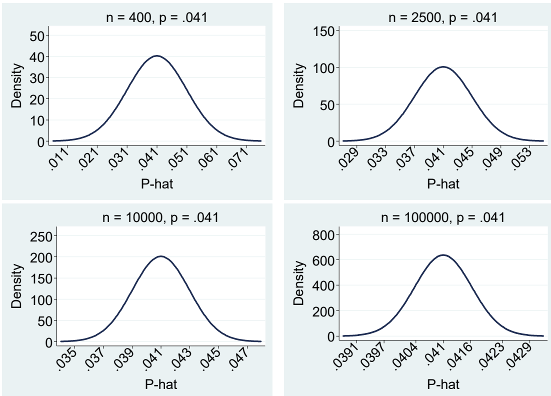




$$P(X > 431) = P\left(Z > \frac{431 - 410}{\sqrt{10{,}000 * 0.041 * 0.956}}\right) = P(Z > 1.06) = 0.14$$

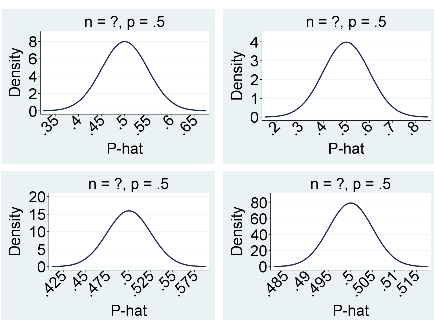
$$P(\hat{P} > 0.0431) = P\left(Z > \frac{0.0431 - 0.041}{\sqrt{\frac{0.041(1-0.041)}{10{,}000}}}\right) = P(Z > 1.06) = 0.14$$

24



As the sample size goes up, sampling error goes down

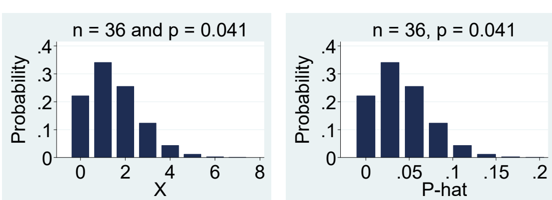
Sampling Error and Sample Size



Recall that $SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}}$ and recall the Empirical (68-95-99.7) Rule.

Which of these corresponds to a sample size of $n = 400$?

What if Normal Approx. Poor?



$$P\left(\hat{P} \geq \frac{2}{36}\right) = 1 - P(X=0) - P(X=1)$$

$$= 1 - \frac{36!}{0!\,36!}0.041^0(1-0.041)^{36} - \frac{36!}{1!\,35!}0.041^1(1-0.041)^{35}$$

$$= 1 - 0.2215 - 0.3410 = 0.4375$$