6. CHANCE VARIABILITY (FPP, Ch 16,17,18, etc.)
6.1 SAMPLING VARIABILITY

- **When we toss a (fair) coin: we do not get the same result every time.**

- **When we toss a fair coin 100 times, we probably won't get exactly 50 heads, but we will get about 50% heads.**

- **When we toss a fair coin 1000 times, we would be very surprised to get EXACTLY 500 heads. But we will get very close to 50% heads.**

- **When we take a random sample from a population, we do not get the same sample every time. The results will be a bit different.**

- **If we (could) repeat a randomized controlled experiment: different subjects would be randomized to treatment/control. The results will be a bit different.**

- **If have someone measure our own height to 0.01 inches, it will not be the same every time. Part of this is measurement error, part may be true variation – in the way we stand, the time of day, ...**

- **NOT ALL VARIATION IS ERROR.**

Blank page for your notes

6.2 THE LAW OF AVERAGES: FPP Ch 16

- Back to the fair coin, tossed many times.
  On average, we will get 50% heads.

number of	within 10 of 50%		40% to 60% of tosses	
tosses	number	chance	number	chance
1	0 or 1	100%	—	—-
10	0 to 10	100%	4 to 6	66%
50	15 to 35	99.7%	20 to 30	88%
100	40 to 60	96%	40 to 60	96%
1,000	490 to 510	49%	400 to 600	∼ 100%
10,000	4990 to 5010	16%	4000 to 6000	100%

- As number of tosses goes up: the chance of being
  within a given number of expected – goes down
  within a given percent of expected – goes up

- FPP calls the difference between observed and expected the chance error
  The chance error in number of heads goes up
  The chance error in proportion of heads goes down

- The LAW OF AVERAGES says
  The chance error in proportion of heads goes down

- The LAW OF AVERAGES does NOT say
  because we had more heads, the chance of getting a head goes down

Blank page for your notes

6.3 BOX MODELS: FPP Ch 16

- **Many chance models can be most easily thought of as drawing tickets repeatedly from a box.**
  Put the tickets back each time!!

- **Tossing a fair coin: counting number of heads**
  2 tickets: – we draw with replacement
  one labeled "0" (tails), one labeled "1" (heads)
  Number of tosses = number of draws.
  Number of heads = sum of values on the tickets.

- **Tossing an unfair coin; suppose coin gives 20% heads**
  box of 100 tickets: 20 with "1" and 80 with "0"
  box of 10 tickets: 2 with "1" and 8 with "0"
  box of 5 tickets: 1 with "1" and 4 with "0".
  Number of heads = sum of values on the tickets.

- **Same unfair coin:**
  suppose a head wins me \$5: tail loses \$1.
  box of 5 tickets: 1 with "5", 4 with "-1"
  Number of tosses = number of draws from box.
  Total winnings = sum of values on the tickets.

- **We don't have to be tossing coins:**
  Chance baby is a boy: just about 50%
  Chance of correct answer in 5 choices: 20%
  Colors of M-and-M's: 10% orange, 30% green,
  50% yellow, 10% brown: Box of 10 tickets
  1 orange, 3 green, 5 yellow, 1 brown.

Blank page for your notes

6.4 PROBABILITY HISTOGRAMS: FPP Ch 18.2

- A histogram of sample values represents the proportions by areas: the total area is 100%.

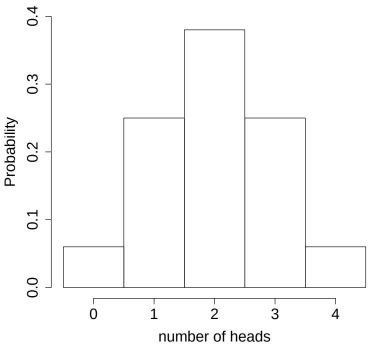
- A probability histogram represents the chances of outcomes by areas: the total area is 100%.

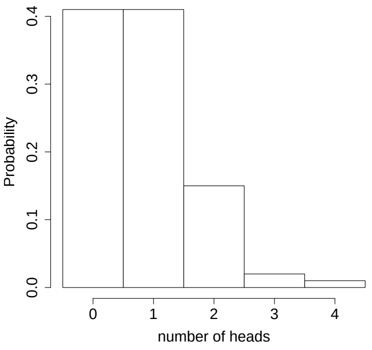
- Toss a fair coin 4 times: we could get 0, 1, 2, 3, 4 heads. The chances of these 5 possibilities are 6.25%, 25%, 37.5%, 25%, and 6.25%.

- What does this mean?
  Proportion of times something happens in many, many repetitions.

- If the coin has only 20% chance of showing heads, the chances are 41%, 41%, 15%, 2%, 1%.

- The probability histograms are





Blank page for your notes

6.5 EXPECTED VALUES AND STANDARD ERRORS
FPP Ch 17

- Suppose the values on the tickets are quantitative.

- Suppose a large number of people each make one draw from the box, with replacement. On average, the value of their ticket is the average of the ticket values in the box: the box average.

- This "on average" value is the expected value

- Suppose we make some number of draws from the box, with replacement, and add them up
expected value = (number of draws) × (box-average)

- But probably we will not get exactly the expected value – there is chance variation.

- The difference of our value from the expected value is the chance error.
How big do we expect the chance error to be?
Answer: the standard error (SE)

- For a sum of draws from a box:
$\text{SE} = \sqrt{\text{number of draws}} \times$ (SD of box).

- 100 times as many draws: SE multiplied by only 10
- If SD of box is large, SE of sum will be large.

Blank page for your notes

6.6 CHANCE VARIATION IN COUNTS

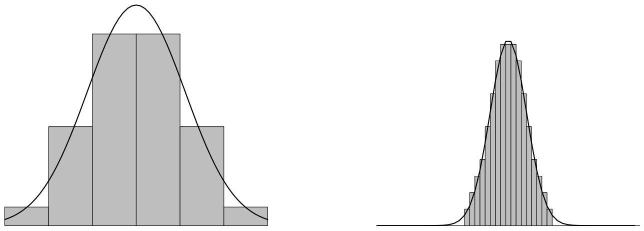
- **For counts: we have only two types of tickets ”1” and ”0”.**
- **For a box with only ”1” and ”0” tickets**
  SD of box $= \sqrt{\text{fraction of ”1”} \times \text{fraction of ”0”}}$
- **For the fair coin: one ”1” and one ”0”**
  On single draw: expectation = 0.5
  SE = (SD of BOX) $= \sqrt{(1/2) \times (1/2)} = 1/2 = 0.5$
- **For the 20% heads coin: one ”1” and four ”0”**
  On single draw: expectation = 0.2
  SE = (SD of BOX) $= \sqrt{0.2 \times 0.8} = 0.4$

		Fair coin			20% heads coin	
number	exp-	SE	expected	exp-	SE	expected
of tosses	ected		± 2 SE	ected		± 2 SE
1	—	0.5	—	—	0.4	—-
10	5	1.58	2 to 8	2	1.26	0 to 5
50	25	3.53	18 to 32	10	2.82	5 to 15
100	50	5.00	40 to 60	20	4.00	12 to 28
1,000	500	15.8	469 to 531	200	12.6	175 to 225
10,000	5000	50.0	4900 to 5100	2000	40.0	1920 to 2080

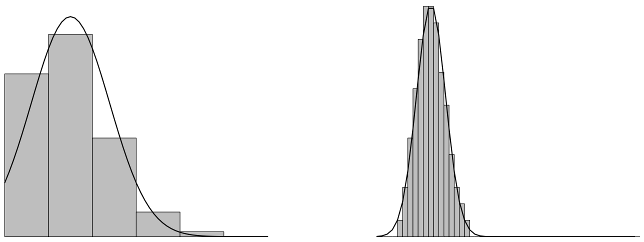
Blank page for your notes

6.7 THE NORMAL APPROXIMATION FOR PROBABILITIES

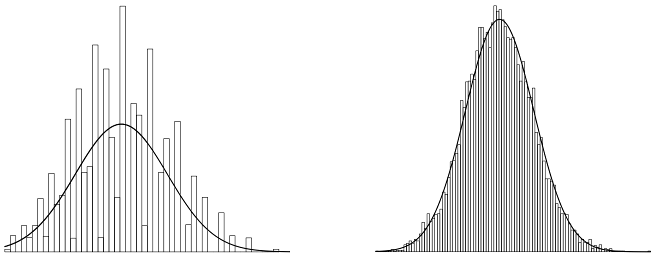
- **Toss a fair coin: 5 times, 50 times**



- **Toss a coin with chance 20% heads: 5, 50 times**



- **Box with 4 tickets, values 0, 2, 5, 10:**
  sum 5 draws, 50 draws



Blank page for your notes

6.8 THE CENTRAL LIMIT THEOREM: FPP Ch 18

- **Fair coin: box mean = 0.5, box SD = 0.5**
  sum of 5 draws, expected = 2.5, SE = 1.12
  sum of 50 draws, expected = 25, SE = 3.54

- **20% heads coin: box mean = 0.2, box SD = 0.4**
  sum of 5 draws, expected = 1.0, SE = 0.89
  sum of 50 draws, expected = 10, SE = 2.83

- **4-ticket box: values 0, 2, 5, 10:**
  box mean 4.25, SD=3.77
  sum of 5 draws, expected = 21.25, SE = 8.42
  sum of 50 draws, expected = 212.5, SE = 26.63

- **As number of draws increases, the probability histogram always gets closer to normal shape, with**
  normal distribution mean = expected
  normal distribution SD = SE

- **So we can <u>standardize</u> the values:**
  Recall for population histograms: (value-mean)/SD
  Now for probability histograms: (value - expected)/SE

- **The probability histogram for the standardized value of the sum gets close to the <u>standard normal curve</u>.**

- **So we can use the table on FPP, P.A105.**

Blank page for your notes

6.9 SAMPLING POPULATIONS and DRAWS FROM BOXES: FPP Ch 20

• When we sample subjects from a population, we observe a value or characteristic associated with each:
height of an individual
income of a household
miles per year driven by a car (or driver)
whether voter will vote ”D” or ”R”
whether vehicle is SUV (yes/no questions)

• When we do repeated draws from a box, we observe the “value” on the ticket.

• From a population we sample without replacement.
From a box we draw with replacement.
But for a large population it makes no difference

• Population histograms give us the distribution of incomes in the population: for example, the percentage in each \$10K interval.

• Now make a box with 100 tickets, and label the right proportion with each \$10K interval. For example, if 8% of household have incomes \$50K to \$59K, label 8 of the 100 tickets ”\$50K to \$59K”.

• Repeated draws for the box with replacement is just like sampling from the population. The probability histogram for the box is like the population or sample histogram (in intervals of \$10K).

• We can use our box models to find out what samples from the population will look like – means, SD, etc.

Blank page for your notes

6.10 CHANCE VARIATION IN PROPORTIONS: FPP Ch 21

- In a population some percentage will have a given characteristic of interest. For example, 55% will vote Democrat.

- We take a sample of 3000 (say) registered voters.

- Taking a sample size n=3000 from a large population (without replacement) is almost like taking n=3000 draws from a box.

- A population with 55% people who will vote Democrat, is like a box, with 100 tickets, 55% marked ”D” or ”1”, 45% marked ”R” or ”0”.

- Or, use just 20 tickets: 11 with ”1”, 9 with ”0”.

- In 6.2, the law of averages showed us that
  while chance error in count got larger,
  the chance error in proportion got smaller,
  as the number of draws gets larger.

- The SE for the proportion of 1's in n draws =
  $\sqrt{n} \times$ (SD of box)/n = (SD of box)/$\sqrt{n}$.

- For our example: SD of box = $\sqrt{0.55 \times 0.45}$=0.497
  For 3000 draws; SE of proportion = 0.497/$\sqrt{3000}$
  = 0.009. Or just under 1%

- So, in sampling: we expect 55% of ”D” tickets and SE is just about 1% if we sample 3000 voters.

- The normal distribution works for us as before: 95% of the time our chance error is less than 2 SE, or 2 percentage points.

Blank page for your notes

6.11 CHANCE VARIATION IN AVERAGES: FPP Ch 23

- For draws from a box: sum the ticket values
  expected sum = (number of draws) × (box average)
  SE of sum = $\sqrt{\text{number of draws}}\times$ (box SD)

- Now take average of the ticket values drawn from box:
  expected average = (box average)
  SE of average = SE for sum / (number of draws)
  = (box SD) /$\sqrt{\text{number of draws}}$

- As before: the normal distribution curve can be used to figure the chances for the average.
  In 95% of repetitions, average is within 2 SE of box average. In 68% of repetitions, average is within 1 SE of box average

- We take a sample from a population to find out about characteristics of the population – for example, average household income.

- Example: sample 1000 households from 100,000 in city
  sample average should be about population (box) average: it differs by the chance error
  SE of average = (SD of box)/$\sqrt{1000}$
  SD of sample should be about the population SD

- As before: the normal distribution curve can be used to figure the chances for the average.
  95% of repetitions, sample average is within 2 SE of population average
  68% of repetitions, sample average is within 1 SE of population average

Blank page for your notes