4. Probability Distributions

Probability: With random sampling or a randomized experiment, the *probability* an observation takes a particular value is the proportion of times that outcome would occur in a long sequence of observations.

Usually corresponds to a *population proportion* (and thus falls between 0 and 1) for some real or conceptual population.

Basic probability rules

Let A, B denotes possible outcomes

- P(not A) = 1 – P(A)

- For distinct possible outcomes A and B,

  P(A or B) = P(A) + P(B)

- P(A and B) = P(A)P(B given A)

- For “independent” outcomes, P(B given A) = P(B), so P(A and B) = P(A)P(B).

(data from 2006 GSS)

	Happiness			
Income	Very	Pretty	Not too	Total
Above Aver.	272	294	49	615
Average	454	835	131	1420
Below Aver.	185	527	208	920
Total	911	1656	388	2955

Let A = average income, B = very happy

P(A) estimated by 1420/2955 = 0.481 (a “marginal probability”),
P(not A) = 1 – P(A) = 0.519

P(B given A) estimated by 454/1420 = 0.320
(a “conditional probability”)

P(A and B) = P(A)P(B given A) est. by 0.481(0.320) = 0.154
(which equals 454/2955, a “joint probability”)

B1: randomly selected person is very happy

B2: second randomly selected person is very happy

P(B1), P(B2) estimated by 911/2955 = 0.308

P(B1 and B2) = P(B1)P(B2) estimated by

(0.308)(0.308) = 0.095

If instead B2 refers to partner of person for B1, B1 and B2 probably not independent and this formula is inappropriate

Probability distribution of a variable

Lists the possible outcomes for the “random variable” and their probabilities

Discrete variable: Assign probabilities P(y) to individual values *y*, with

$$0 \leq P(y) \leq 1, \quad \Sigma P(y) = 1$$

Example: Randomly sample 3 people and ask whether they favor (F) or oppose (O) a public health care system

y = number who “favor” (0, 1, 2, or 3)

For possible samples of size *n* = 3,

Sample	*y*	Sample	*y*
(O, O, O)	0	(O, F, F)	2
(O, O, F)	1	(F, O, F)	2
(O, F, O)	1	(F, F, O)	2
(F, O, O)	1	(F, F, F)	3

If population equally split between F and O, these eight samples are equally likely and probability distribution of *y* is

y	*P(y)*
0	1/8
1	3/8
2	3/8
3	1/8

(special case of “binomial distribution,” introduced in Chap. 6). In practice, probability distributions are often estimated from sample data, and then have the form of frequency distributions

Example: GSS results on *y* = number of people you knew personally who committed suicide in past 12 months (variable “suiknew”).

Estimated probability distribution is

y	*P(y)*
0	.895
1	.084
2	.015
3	.006

Like frequency distributions, probability distributions have descriptive measures, such as mean and standard deviation

- Mean (*expected value*) -

$$\mu = E(Y) = \sum yP(y)$$

μ = 0(0.895) + 1(0.084) + 2(0.015) + 3 (0.006) = 0.13

represents a “long run average outcome”

(median = mode = 0)

Standard Deviation - Measure of the “typical” distance of an outcome from the mean, denoted by σ

$$\sigma = \sqrt{\Sigma(y-\mu)^2 P(y)}$$

(We won’t need to calculate this formula.)

If a distribution is approximately bell-shaped, then:

- all or nearly all the distribution falls between
  μ - 3σ and μ + 3σ

- Probability about 0.68 falls between
  μ - σ and μ + σ

Example: From result later in chapter, if *n* people are randomly selected from population with proportion π favoring public health care (1-π, Oppose), then

y = *number* in sample who favor it

has a bell-shaped probability distribution with

$$\mu = E(y) = n\pi, \ \sigma = \sqrt{n\pi(1-\pi)}$$

e.g, with *n = 1000, π = 0.50,   get μ = 500, σ = 16*

Nearly all the distribution falls between about

500 – 3(16) = 452 and 500 + 3(16) = 548

i.e., almost certainly between about 45% and 55% of a sample will say they favor public health care

Continuous variables: Probabilities assigned to intervals of numbers

Ex. When *y* takes lots of values, as in last example, it is continuous for practical purposes. Then, if probability distribution is approx. bell-shaped,

$$P(\mu - \sigma \leq y \leq \mu + \sigma) \approx 0.68, \; P(\mu - 2\sigma \leq y \leq \mu + 2\sigma) \approx 0.95$$

In previous example, $P(\mu - \sigma \leq y \leq \mu + \sigma) = P(484 \leq y \leq 516) \approx 0.68$

Most important probability distribution for continuous variables is the **normal distribution**

Normal distribution

- Symmetric, bell-shaped (formula in Exercise 4.56)
- Characterized by mean (μ) and standard deviation (σ), representing center and spread
- Probability within any particular number of standard deviations of μ is same for all normal distributions
- An individual observation from an approximately normal distribution has probability
  - ➢ 0.68 of falling within 1 standard deviation of mean
  - ➢ 0.95 of falling within 2 standard deviations
  - ➢ 0.997 of falling within 3 standard deviations

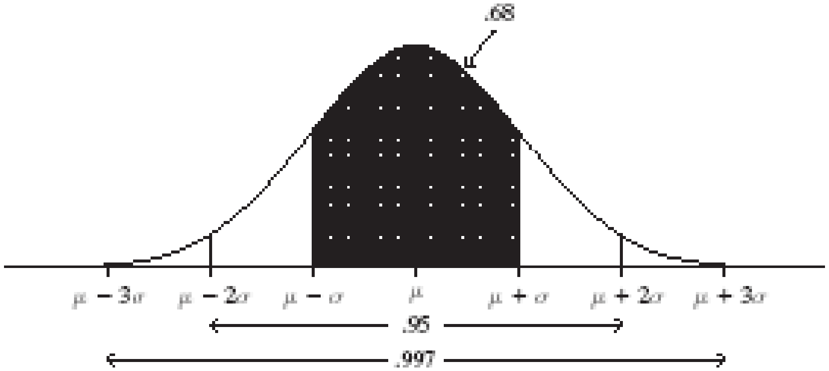

.68
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ
.95
.997

Table A (inside back cover of text, and in packet of tables at course home page) gives probability in right tail above μ + zσ for various values of z.

Second Decimal Place of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
		…								
		….								
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
…….										
……..										

Example: What is probability falling between μ - 1.50σ and μ + 1.50σ ?

- z = 1.50 has right tail probability = 0.0668
- Left tail probability = 0.0668 by symmetry
- Two-tail probability = 2(0.0668) = 0.1336
- Probability within μ - 1.50σ and μ + 1.50σ is

  1 – 0.1336 = 0.87

Example: z = 2.0 gives

two-tail prob. = 2(0.0228) = 0.046,

probability within $\mu \pm 2\sigma$ is 1 - 0.046 = 0.954

Example: What z-score corresponds to 99th percentile (i.e., $\mu + z\sigma$ = 99th percentile)?

- Right tail probability = 0.01 has $z = 2.33$
- 99% falls below $\mu + 2.33\sigma$

If IQ has $\mu = 100$, $\sigma = 16$, then 99th percentile is

$$\mu + 2.33\sigma = 100 + 2.33(16) = 137$$

Note: $\mu - 2.33\sigma = 100 - 2.33(16) = 63$ is 1st percentile

0.98 = probability that IQ falls between 63 and 137

Example: What is z so that $\mu \pm z\sigma$ encloses exactly 95% of normal curve?

- Total probability in two tails = 0.05
- Probability in right tail = 0.05/2 = 0.025
- $z = 1.96$

$\mu \pm 1.96\sigma$ contains probability 0.950

($\mu \pm 2\sigma$ contains probability 0.954)

Exercise: Try this for 99%, 90%

(should get 2.58, 1.64)

Example: Minnesota Multiphasic Personality Inventory (MMPI), based on responses to 500 true/false questions, provides scores for several scales (e.g., depression, anxiety, substance abuse), with

$$\mu = 50, \sigma = 10.$$

If distribution is normal and score of ≥ 65 is considered abnormally high, what percentage is this?

- $z = (65 - 50)/10 = 1.50$
- Right tail probability = 0.067 (less than 7%)

Notes about z-scores

- z-score represents *number of standard deviations* that a value falls from mean of dist.

- A value y is $z = (y - \mu)/\sigma$ standard deviations from μ

Example: $y = 65$, $\mu = 50$, $\sigma = 10$

$$z = (y - \mu)/\sigma = (65 – 50)/10 = 1.5$$

- The z-score is negative when y falls below μ (e.g., y = 35 has $z = -1.5$)

- The **standard normal distribution** is the normal distribution with $\mu = 0$, $\sigma = 1$

For that distribution, $z = (y - \mu)/\sigma = (y - 0)/1 = y$

i.e., original score = z-score

$\mu + z\sigma = 0 + z(1) = z$

(we use standard normal for statistical inference starting in Chapter 6, where certain statistics are scaled to have a standard normal distribution)

- *Why is normal dist. important*?

We’ll learn today that if different studies take random samples and calculate a statistic (e.g. sample mean) to estimate a parameter (e.g. population mean), the collection of statistic values from those studies usually has approximately a normal distribution. (So?)

A ***sampling distribution*** lists the possible values of a statistic (e.g., sample mean or sample proportion) and their probabilities

Example: y = 1 if favor health care plan

y = 0 if oppose

For possible samples of size *n* = 3, consider sample mean

Sample	Mean	Sample	Mean
(1, 1, 1)	1.0	(1, 0, 0)	1/3
(1, 1, 0)	2/3	(0, 1, 0)	1/3
(1, 0, 1)	2/3	(0, 0, 1)	1/3
(0, 1, 1)	2/3	(0, 0, 0)	0

For binary data (0, 1), *sample mean equals sample proportion* of “1” cases. For population,

$$\mu = \sum yP(y)\text{=0P(0)+1P(1)=P(1)}$$

is population proportion of “1” cases (e.g., favoring public health care)

How close is sample mean to population mean μ?

To answer this, we must be able to answer,
“What is the probability distribution of the sample mean?”

Sampling distribution of a statistic is the probability distribution for the possible values of the statistic

Ex. Suppose P(0) = P(1) = ½. For random sample of size *n* = 3, each of 8 possible samples is equally likely. Sampling distribution of *sample proportion* is

Sample proportion	Probability	
0	1/8	
1/3	3/8	
2/3	3/8	
1	1/8	(Try for *n* = 4)

$\bar{y}$

Sampling distribution of sample mean

- $\bar{y}$ is a variable, its value varying from sample to sample about the population mean μ
- Standard deviation of sampling distribution of $\bar{y}$ is called the **standard error of** $\bar{y}$
- For random sampling, the sampling distribution of $\bar{y}$ has mean μ and standard error

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

- **Example**: For binary data (*y* =1 or 0) with P(Y=1) = π (with 0 < π < 1), can show that

$\sigma = \sqrt{\pi(1-\pi)}$ (Exercise 4.55b, and special case of earlier formula on p. 11 of these notes with *n* = 1)

When π = 0.50, σ = 0.50, and standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{n}}$$

n	standard error
3	.289
100	.050
200	.035
1000	.016

- Note standard error goes down as *n* goes up

  (i.e., $\bar{y}$ tends to fall closer to μ)

- With *n* = 1000, standard error = 0.016, so if the sampling distribution is bell-shaped, with very high probability the sample proportion falls within 3(0.016) = 0.05 of population prop of 0.50

  (i.e., between about 0.45 and 0.55)

Number of times y = 1 *(*i.e., number of people in favor) is 1000×(proportion), so that the "count" variable has

mean = 1000(0.50) = 500

std. dev. 1000(0.016) = 16 (as in earlier ex. on p. 11)

- Practical implication: This chapter presents *theoretical* results about spread (and shape) of *sampling distributions,* but it implies how different studies on the same topic can vary from study to study *in practice* (and therefore how *precise* any one study will tend to be)

Ex. You plan to sample 200 people to estimate the *population* proportion supporting a proposed health care plan. Other people could be doing the same thing. How will the results *vary among studies* (and how *precise* are your results)?

The sampling distribution of the *sample* proportion in favor of the health care plan has *standard error* describing the likely variability from study to study.

Ex. Many studies each take sample of *n = 200* to estimate population proportion

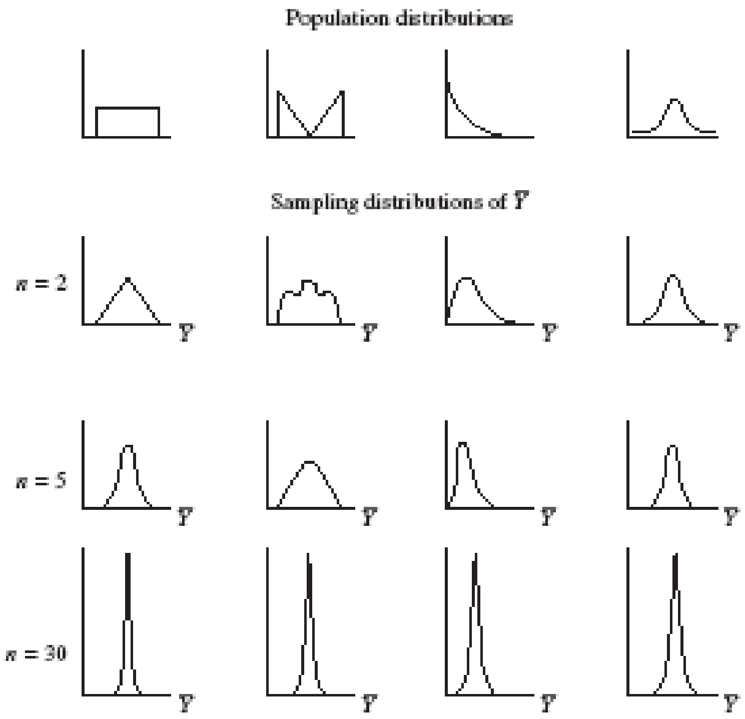
- Flipping a coin 200 times simulates the process when the population proportion = 0.50.
- In theory, we've seen the sample proportion should vary from study to study (i.e., from student to student) around 0.50 with a standard error of 0.035

- Empirical evidence: I took the data you generated, and I calculated that the set of all the sample proportions (0.515 = 103/200, 0.470 = 94/200, etc.) had a mean of 0.488 and a standard deviation of 0.028. (OK, I cheated and deleted an outlier of 0.67)
- Shape?   Roughly bell-shaped. Why?

Central Limit Theorem: For random sampling with "large" *n,* the sampling dist. of the sample mean $\bar{y}$ is approximately a normal distribution

- Approximate normality applies *no matter what the shape* of the population dist. (Figure p. 93, next page)
- How "large" *n* needs to be depends on skew of population distribution, but usually *n* ≥ *30* sufficient
- Can be verified empirically, by simulating with "sampling distribution" applet at

www.prenhall.com/agresti


Population distributions
Sampling distributions of $\bar{Y}$
n = 2
n = 5
n = 30
$\bar{Y}$

Example: Random sample of 100 students selected to estimate proportion who have participated in "binge drinking." Find probability the sample proportion falls within 0.04 of population proportion, if that population proportion = 0.30 (i.e., between 0.26 and 0.34)

y = 1, yes   y = 0, no

μ = π = 0.30, $\sigma = \sqrt{\pi(1-\pi)} = \sqrt{(0.3)(0.7)} = 0.458$

By CLT, sampling distribution of sample mean (which is the proportion "yes") is approx. normal with mean 0.30,

standard error =

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.458}{\sqrt{n}} = \frac{0.458}{\sqrt{100}} = 0.0458$$

- 0.26 has z-score z = (0.26 - 0.30)/0.0458 = -0.87
- 0.34 has z-score z = (0.34 - 0.30)/0.0458 = 0.87
- P(sample mean ≥ 0.34) = 0.19
- P(sample mean ≤ 0.26) = 0.19
- P(0.26 ≤ sample mean ≤ 0.34) = 1 – 2(0.19) = 0.62

The probability is 0.62 that the sample proportion will fall within 0.04 of the population proportion

Ex. : Coin flipping, n = 200 per student

- If the probability of a head = 0.50, then the sample proportion of heads in 200 flips should vary from student to student according to a *normal distribution* (Did it?) with

  mean = 0.50

  standard error = 0.035 (how?)

It would be very unusual for the proportion of heads to be below 0.40 or above 0.60 (why?)

How would the interval of likely values (0.40, 0.60) change as *n* gets larger? (e.g., *n* = 1000 in a poll)

Don’t be “fooled by randomness”

- We’ve seen that some things are quite predictable (e.g., how close a sample mean falls to a population mean, for a given *n*)

- But, in the short term, randomness is not as “regular” as you might expect (I can usually predict who “faked” coin flipping.)

- In 200 flips of a fair coin,

  P(longest streak of consecutive H’s < 5) = 0.03

  The probability distribution of longest streak has μ = 7

Implications: sports (win/loss, individual success/failure)
stock market up or down from day to day, ….

Some Summary Comments

- Consequence of CLT: When the value of a variable is a result of averaging many individual influences, no one dominating, the distribution is approx. normal (e.g., IQ, blood pressure)

- In practice, we don't know μ, but we can use spread of sampling distribution as basis of inference for unknown parameter value

(we'll see how in next two chapters)

- We have now discussed three *types* of distributions:

- **Population** distribution – described by parameters such as μ, σ (usually unknown)

- **Sample data** distribution – described by sample statistics such as

  sample mean $\bar{y}$, standard deviation *s*

- **Sampling** distribution – probability distribution for possible values of a sample statistic; determines probability that statistic falls within certain distance of population parameter

(graphic showing differences)

Ex. (categorical): Poll about health care

Statistic = sample proportion favoring a proposed health care plan

What is (1) population distribution, (2) sample distribution, (3) sampling distribution?

Ex. (quantitative): Experiment about impact of cell-phone use on reaction times

Statistic = sample mean reaction time

What is (1) population distribution, (2) sample distribution, (3) sampling distribution?

By the Central Limit Theorem (multiple choice)

- All variables have approximately normal sample data distributions if a random sample has at least 30 observations
- Population distributions are normal whenever the population size is large (at least about 30)
- For large random samples, the sampling distribution of the sample mean is approximately normal, regardless of the shape of the population distribution
- The sampling distribution looks more like the population distribution as the sample size increases
- All of the above