

## Producing Data/Data Collection

Without serious care/thought here, all is lost ... no amount of clever post-processing of useless data will make it informative. GIGO

Chapter 3 of MMD&S is an elementary discussion of some aspects of data collection. The chapter draws a sharp contrast between "observational studies" and "experiments." The distinction is in the degree to which an investigator manipulates the system under study, as opposed to simply passively observing it. In truth, no real study is probably ever purely observational or purely experimental, and the two ideals set up by MMD&S should probably be thought of as idealized endpoints on a continuum of study types with varying degrees of investigator intervention/manipulation.

1

"Observational Studies"

"Experimental Studies"



Figure 1: A Continuum

Two essential points made in Ch 3 of MMD&S about real world data collection are

1. one essentially never has "all possible data,"

2

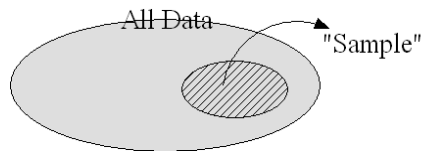


Figure 2: Sampling

2. randomness/chance always affects data that are collected ... often because an investigator purposely injects it into the data collection process.

Video: Against All Odds, Unit #13, Frito-Lay

3

Intentional injection of "randomness" into sample selection/data collection has at least 2 purposes:

1. protection against conscious or unconscious bias (it provides a degree of objectivity),
2. it allows the use of the mathematics of probability (the mathematical description of "randomness") in data analysis.

*Examples (of intentional use of "randomness" in data collection)*

*Experimental Contexts*

4

- Aspirin study- half of the subjects were chosen "at random" to get aspirin treatment
- Pizza taste test- the order of presentation of cheese and double cheese was "randomly determined"

*An Observational Context*

- 10,000 case files, 100 to be selected "at random" for quality auditing of paperwork

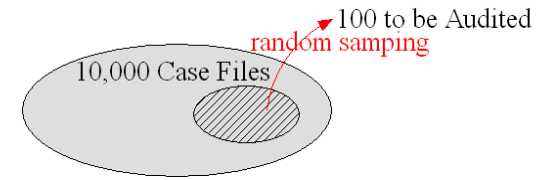


Figure 3: Random Sampling of Case Files

Exactly what is "random sampling"? It's anything/any method that is conceptually equivalent to "drawing slips of paper from a hat." More formally:

**A simple random sample of  $n$  items from  $N$  items** is one chosen in such a way that *a priori* every group of  $n$  is "equally likely" to compose the sample.

How does one obtain a simple random sample?

- using tables of random digits (not common these days)
- using some mechanical randomizing device (also not so common)
- using a computer-implemented numerical algorithm that generates "pseudo-random" digits

*Example* Using JMP to pick  $n = 100$  from  $N = 10,000$  case numbers

The most basic mental model of data collection is that one is selecting values as a simple random sample from a very large population ... usually thought to be so large that there is little practical difference between random sampling

without replacement (simple random sampling) and sampling with replacement. "Randomness" is very close conceptually to "variation." It enters data collection processes both because it is sometimes put there by an investigator, and also because of many unnamed small causes that I don't even try to account for. If it is going to be a part of data collection and analysis we need some basics of it in Stat 328.

## Probability: The Mathematical Description of "Chance" / "Randomness"

This is a mathematical theory, just like geometry was a mathematical theory or system that you studied in high school.

## Basic Terminology of Probability (Set Theory)

A specification of all possible outcomes of a "chance" scenario" is called a **sample space**, and is typically symbolized as  $\mathcal{S}$ . (This is the universe or universal set of set theory jargon.)

*Example* Weekly "Powerball" drawing

The powerball can have any number on it, from 1 to 42. So

$$\mathcal{S} = \{1, 2, 3, \dots, 41, 42\}$$

A group of *some* of all possible outcomes is called an **event**. Capital letters near the beginning of the alphabet are used to stand for events. (In "set theory" jargon, an event is a subset of the sample space/universal set,  $\mathcal{S}$ .)

9

*Example* Weekly "Powerball" drawing

$$\begin{aligned} A &= \{2, 4, 6, \dots, 40, 42\} \\ &= \text{"the powerball is even"} \end{aligned}$$

and

$$\begin{aligned} B &= \{1, 2, 3, 4, 5\} \\ &= \text{"the powerball is 5 or less"} \end{aligned}$$

10

The event made up of all outcomes in one or both of  $A, B$  is called

$$A \text{ or } B \quad (= A \cup B)$$

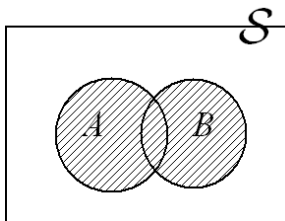


Figure 4:  $A \text{ or } B$

11

The event made up of all outcomes common to  $A, B$  is called

$$A \text{ and } B \quad (= A \cap B)$$

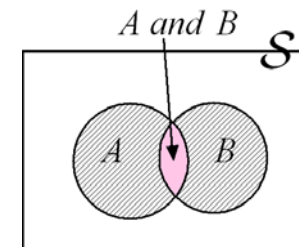


Figure 5:  $A \text{ and } B$

12

The event made up of all outcomes in  $\mathcal{S}$  but not in  $A$  is called

$$\text{not } A \quad (= A^c = \bar{A})$$

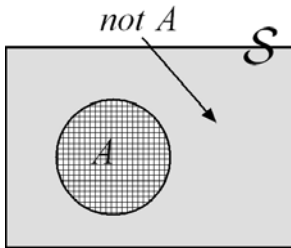


Figure 6: *not A*

Probability theory isn't just set theory ... rather, it is about assigning measures of "likelihood" or "chance" to events in a mathematically consistent/rational/

coherent way. That requires axioms or "rules of the game" ... just like high school geometry is developed from a few basic postulates/axioms.

### Axioms/Ground Rules for Probability

A **probability model** is an assignment of "probability values" (likelihoods)  $P$  to events that satisfies

1.  $0 \leq P(A) \leq 1$  for all events  $A$
2.  $P(\mathcal{S}) = 1$

3.  $P(\text{not } A) = 1 - P(A)$  for all events  $A$

4. for any two disjoint (non-overlapping) events

$$P(A \text{ or } B) = P(A) + P(B)$$

Any system that satisfies these axioms is a mathematically valid probability model. Whether or not that model usefully describes reality is an entirely different question ... one that can only be answered empirically/on the basis of data.

*Example* Powerball Saturday

A mathematically valid assignment of probabilities is

$$P(A) = \frac{\# \text{ of outcomes in } A}{\# \text{ of outcomes in } \mathcal{S}} = \frac{\# \text{ of outcomes in } A}{42}$$

Using this model, if

$$B = \{1, 2, 3, 4, 5\}$$

then

$$P(B) = \frac{5}{42}$$

And if

$$C = \{20\}$$

then

$$P(C) = \frac{1}{42}$$

Further, with  $P(B) = \frac{5}{42}$ , it follows immediately that

$$P(\text{not } B) = 1 - \frac{5}{42} = \frac{37}{42}$$

And using words to describe events rather than using set notation

$$\begin{aligned} P(\text{powerball is less than 10 or larger than 30}) &= \frac{9}{42} + \frac{12}{42} \\ &= \frac{21}{42} = .5 \end{aligned}$$

17

Whether the "equally likely outcomes" model used above for Powerball is a good one or not is an empirical matter. Data at

[http://www.powerball.com/powerball/pb\\_frequency.asp](http://www.powerball.com/powerball/pb_frequency.asp)

actually suggested (through June 2004) that the "20 ball" had come up "too frequently" in the past for this model to be a good one.

*Exercises* Problems 4.15, 4.19, 4.26 of MMD&S

## Random Variables

These are "chance quantities"/values depending upon the outcome of some chance situation. Since the exact value of a random variable can't be predicted,

18

it is important to instead describe its **probability distribution**. This consists of

1. the set of possible values,
2. the probabilities with which those values are taken on.

There are two elementary kinds of distributions used to model random variables, **discrete distributions** and **continuous distributions**. **Discrete** distributions are used where the set of possible values is finite or is perhaps something like the set of integers. **Continuous** models are an idealization where one use a whole interval of numbers for the set of possible values. Different tools are used to specify and describe discrete and continuous distributions. (And strictly speaking, to completely lay out the theory for continuous distributions, one must employ calculus ... something that will not be done in Stat 328.)

19

## Discrete Distributions

It is common (at least while learning the basics of probability) to use capital letters near the beginning of the alphabet to stand for random variables.

The basic tool used to describe discrete random variables is the so called **probability function**,  $p(x)$ . For a random variable  $X$ , this gives for every possible value  $x$ , the corresponding that the realized value is  $x$ . Sometimes one gives a formula for  $p(x)$ . For what we do in Stat 328, it will be adequate to simply think of  $p(x)$  as specified in tabular form.

*Example* Early (pre-Challenger) space shuttle

$X$  = the number of O-ring failures on the solid rocket booster  
(on the "next" launch)

20

$x$	$p(x)$
0	.87
1	.12
2	.01
3	<i>small</i>
4	<i>small</i>
5	<i>small</i>
6	<i>small</i>

This specifies how probabilities for  $X$  are to be assigned. It specifies the "distribution" of (the discrete random variable)  $X$ . (The origin of this table is a completely separate matter ... in fact it comes from a combination of some theoretical calculations and empirical experience before the Challenger launch.) Based on this probability function, one has, for example

$$\begin{aligned}
 P(X \text{ is at least } 1) &= P(X \geq 1) \\
 &= .12 + .01 = .13
 \end{aligned}$$

21

It is useful to summarize a probability function/discrete probability distribution in ways very similar to the ways we summarized a data set/empirical distribution. For example, people make "probability histograms." Note that on a well-made probability histogram, areas correspond to probability.

*Example* Early (pre-Challenger) space shuttle

22

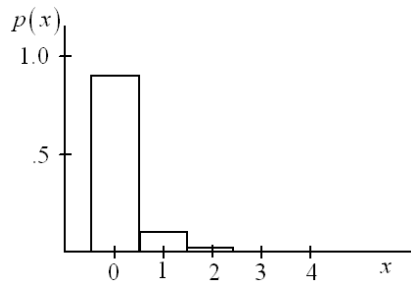


Figure 7: A Probability Histogram for Describing the Number of O-ring Failures

23

There are also notions of "mean" and "standard deviation" for random variables/probability distributions. For discrete cases, these are

$$\mu_X = \sum xp(x)$$

(a weighted average of possible values of  $X$ ) and

$$\sigma_X = \sqrt{\sum (x - \mu_X)^2 p(x)}$$

(the square root of a weighted average squared distance from the realized value of  $X$  to the center of the distribution).

*Example* Early (pre-Challenger) space shuttle

$x$	$p(x)$	$xp(x)$	$(x - \mu_X)^2 p(x)$
0	.87	0(.87)	$(0 - .14)^2 (.87)$
1	.12	1(.12)	$(1 - .14)^2 (.12)$
2	.01	2(.01)	$(2 - .14)^2 (.01)$
		.14	.1404

24

That is,

$$\mu_X = 0(.87) + 1(.12) + 2(.01) = .14 \text{ failures}$$

and

$$\begin{aligned}\sigma_X^2 &= (0 - .14)^2(.87) + (1 - .14)^2(.12) + (2 - .14)^2(.01) \\ &= .1404 \text{ (failures)}^2\end{aligned}$$

so that

$$\sigma_X = .37 \text{ failures}$$

*Exercises* Problems 4.45 and 4.49 of MMD&S

25

## Continuous Distributions

The basic tool for specifying a continuous distribution is a (probability) **density curve**. This is an idealized probability histogram and one simply declares that areas under the density curve are used to specify probabilities. (In general, computation of areas under curves requires the use of calculus. However, for the specific continuous distributions useful in Stat 328, tables are available ... in fact, the standard normal table that we've already used is one such table!)

*Example (of a very simple density curve, where area can be calculated from simple geometry)* Let

$X$  = the next value generated by the "random number" function on my calculator

26

We might model this random variable as continuous. We'd like every interval of numbers between 0 and 1 of a given size to have the same probability assigned to it ... that is, we'd like to have

$$P(0 < X < .2) = P(.17 < X < .37) = .2$$

A density curve that will do this for us is

27

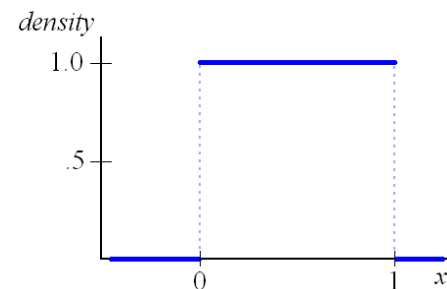


Figure 8: Model for the Output of a "Uniform" Random Number Generator

28

Using this continuous model, one has, for example

$$P(.3 < X < .7) = .7 - .3 = .4$$

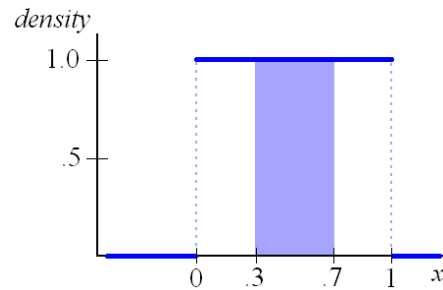


Figure 9:  $P(.3 < X < .7)$

Similarly,

$$P(X > .25) = 1 - .25 = .75$$

and

$$P(X = .64) = 0$$

This last example calculation illustrates the fact that a continuous model assigns 0 probability to any particular possible value ... the area under a curve above a single point is 0. This is slightly counter-intuitive but very convenient. It means that when working with continuous distributions I can be sloppy and not distinguish between  $P(X < a)$  and  $P(X \leq a)$ . (These may be different if  $X$  is discrete.)

Without using the terminology of random variables and density curves, our discussion of normal models actually made use of the "standard normal density curve." I didn't give you the formula for that curve (there is such a specific

formula, but it would not prove informative to you) and we don't use that formula to get areas, but rather the table (that was generated based on the formula).

It is not possible (without using calculus) to say exactly what is meant by mean ( $\mu_X$ ) and standard deviation ( $\sigma_X$ ) for a continuous model. However, you will not go far wrong by thinking

1. a density curve is approximately a probability histogram,
2. the mean for the continuous distribution is approximately the mean for a similar probability histogram (calculation discussed above),

3. the standard deviation for the continuous distribution is approximately the standard deviation for a similar probability histogram (calculation discussed above).

### Sampling Distributions

The simplest use of probability in statistical inference is based on a model that says that  $n$  data values

$$X_1, X_2, \dots, X_n$$

can be thought of as random draws from some fixed population/universe/stable process. Notice that these values are then  $n$  random variables. What is

more, these data values/random variables are typically processed into summary statistics like

$$\bar{X} = \frac{1}{n} \sum X_i \text{ and } S = \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$$

that are then also random variables in their own right that therefore have their own probability distributions! Since random variables like  $\bar{X}$  and  $S$  are summary values for a sample, their probability distributions are typically called **sampling distributions**.

There is a real question as to how one is to understand something as complicated as a sampling distribution, having the minimal probability background of Stat 328 at one's disposal. *Simulations* using statistical software like JMP are one way.

*Example* Consider again the scenario of random numbers generated by a calculator (used above as an example of a simple continuous model). In fact,

now consider hitting the random number button on the calculator twice, and letting

- $X_1$  = the first random number selected
- $X_2$  = the second random number selected

Further, suppose that of interest is the "sum" random variable

$$X_1 + X_2$$

MMD&S on page 247 show the density curve for either one of the variables  $X_1, X_2$ . Some calculus-based theoretical calculations lead to the conclusion that an appropriate density curve for the sum is given on page 248 of MMD&S. That density curve represents the sampling distribution for  $X_1 + X_2$ . An empirical approximation to that theoretical curve can be had by making up two (long) columns in a JMP table based on "uniform random numbers," constructing a third as the (row-at-a-time) sum of the first two, and plotting a histogram for that 3rd row. The histogram will look much like the density curve

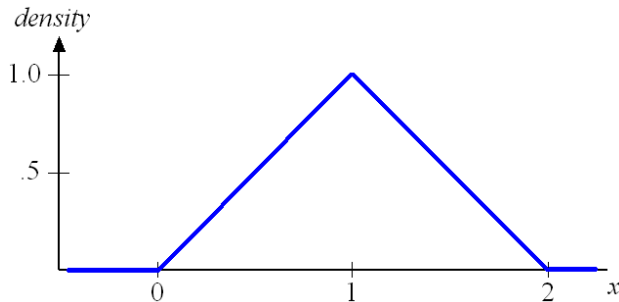


Figure 10: Density for the Sum of Two Uniform Random Numbers

In addition to simulations, there is theory for sampling distributions that can help us understand what to expect. The balance of this session will concern what probability theory promises regarding the **sampling distribution of the sample mean** (the probability distribution of  $\bar{X}$ ).

First, there are the general facts that

$$\mu_{\bar{X}} = \mu$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The first of these says that the probability distribution of the sample mean random variable has the same center as the universe/population/process being sampled. The second says that the spread of probability distribution of the sample mean random variable is related to the spread of the universe/population/process

being sampled, by division of that by the square root of the sample size. (Sample means cluster around the population mean, with a variability that *decreases* as the sample size increases.)

Related to these facts about mean and standard deviation is the so-called **law of large numbers (LLN)** that says that as sample size increases, the probability distribution appropriate for  $\bar{X}$  becomes more and more tightly packed about the population/universe/process mean,  $\mu$ .

Finally, there are facts that imply that in many contexts,  $\bar{X}$  may be treated as if it is at least approximately a normal random variable. (NOTICE that saying that a sample mean  $\bar{X}$  is normal is NOT the same thing as saying that the population from which individuals are being drawn is normal!) That is, there are the facts

1. IF a population/universe/process IS normal, then the probability distribution of  $\bar{X}$  (based on random samples from it) is exactly normal,
2. for essentially any population/universe/process, for "large  $n$ " the probability distribution of  $\bar{X}$  (based on random samples from it) is approximately normal.

Fact 2 is the famous **central limit theorem**. Figure 4.17 on page 293 of MMD&S illustrates a particular case of the CLT. The following illustrates Fact 1 (and the facts that  $\mu_{\bar{X}} = \mu$  and  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ ) pictorially.

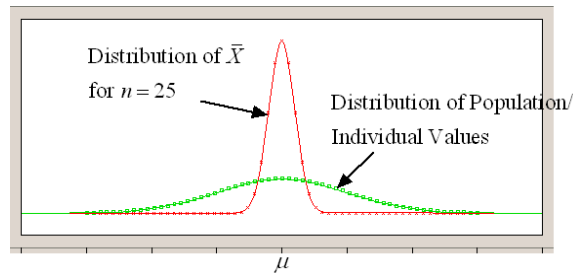


Figure 11: Illustration of the Central Limit Theorem

*Example* Problem 4.86, page 294 MMD&S

*Individual* ACT scores are modeled as normal with  $\mu = 18.6$  and  $\sigma = 5.9$ . (We're going to ignore the fact that only integer ACT scores are possible ... that makes this continuous model less than perfect.)

The chance that a single individual student makes a score of 21 or higher can be pictured as

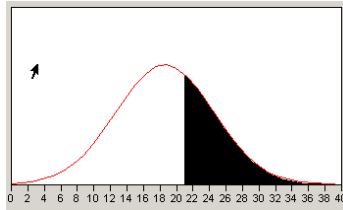


Figure 12: The Chance that a Single Student Scores 21 or Higher

41

Then, since

$$z = \frac{x - \mu}{\sigma} = \frac{21 - 18.6}{5.9} = .41$$

and a direct table look-up with  $z = .41$  gives .6591,

$$\begin{aligned} P(\text{an individual score is 21 or more}) &= P(X \geq 21) \\ &= 1 - .6591 \\ &= .3409 \end{aligned}$$

Then consider the sample average score for a randomly selected group of  $n = 50$  students. For a sample of this size

$$\mu_{\bar{X}} = \mu = 18.6$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5.9}{\sqrt{50}} = .83$$

42

The chance that a the sample average score is 21 or higher can be pictured as

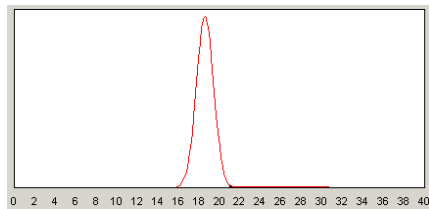


Figure 13: The Probability That an Average of 50 ACT Scores Exceeds 21

43

$$z = \frac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{21 - 18.6}{.83} = 2.89$$

and a direct table look-up with  $z = 2.89$  gives .9981,

$$\begin{aligned} P(\text{the sample mean score is 21 or more}) &= P(\bar{X} \geq 21) \\ &= 1 - .9981 \\ &= .0019 \end{aligned}$$

44