YMS.4e CONTENT REVIEW
(Chapters 4-6)

IV. Designing Studies

A. Sampling and Surveys

1. A census contacts every individual in the population to obtain data; a sample survey collects data from part of a population in order to learn something about the entire population.

2. Bad sampling designs result in bias in different forms

   voluntary response sample – participants choose themselves
   convenience sample – investigators choose to sample those people who are easy to reach

3. Good sampling designs

   simple random sample – a group of *n* individuals chosen from a population in such a way that every set of *n* individuals has an equal chance of being the sample actually chosen; use a random number table or **randint** on the calculator
   stratified random sample – divide the population into groups (strata) of similar individuals (by some chosen category) then choose a simple random sample from each of the groups
   cluster sampling- divide the population into groups (clusters); randomly select some of these clusters; all individuals in chosen clusters are included in the sample.

4. Sampling errors

   Biased sampling design- the design systematically favors certain outcomes or responses
   Undercoverage- when some groups of the population are left out, often because a complete list of the population from which the sample was chosen (sampling frame) was not accurate or available.

5. Nonsampling errors

   nonresponse – when an individual appropriately chosen for the sample cannot or does not respond
   response bias – when an individual does not answer a question truthfully, e.g. a question about previous drug use may not be answered accurately
   wording of questions – questions are worded to elicit a particular response, e.g. One of the Ten Commandments states, "Thou shalt not kill." Do you favor the death penalty?

B. Experiments

1. An observational study observes individuals in a population or sample, measures variables of interest, but does not in any way assign treatments or influence responses

2. An experiment deliberately imposes some treatment on individuals (experimental units or subjects) in order to observe response. *Can* give evidence for causation *if* well designed with a control group. 3 necessities:

Control – for lurking variables by assigning units to groups that do not get the treatment

Lurking variables (variables not identified or considered) may explain a relationship between the explanatory and response variables by either confounding (a third variable affects the response variable only) or by common response (a third variable affects both the explanatory and response variables

Randomize – use simple random sampling to assign units to treatments/control groups
Replicate – use the same treatment on many units to reduce the variation due to chance

3. The "best" experiments are double blind – neither the investigators nor the subjects know which treatments are being used on which subjects. Placebos are often used.

4. Designs

Between groups (independent samples)- sometimes uses blocking where subjects are grouped before the experiment based on a particular characteristic or set of characteristics, then simple random samples are taken within each block.

Within groups (repeated measures)

Matched pairs

C. Using Studies Wisely

1. Inference about the population requires that the individuals in a study be randomly selected

2. Correlational studies *can* provide evidence of causation but it’s tricky

3. Do not automatically accept a study is true without analysis

V. Probability: What Are The Chances

A. Randomness, Probability and Simulation

1. Probability only refers to “the long run” (law of large numbers) never short run

2. A probability is a number between 0 and 1

3. Simulations can be used to determine probabilities

B. Probability Rules

1. All probabilities for one event must sum to 1

2. $P(A^C) = 1 - P(A)$ where A^C is the complement of A

3. *Mutually Exclusive* (*disjoint*) *Events*– events which cannot occur at the same time; mutually exclusive events ALWAYS have an effect on each other so they can never be independent.

4. If $P(A+B) = 0$ then A and B are mutually exclusive and:

$$P(A\ or\ B) = P(A) + P(B) \longrightarrow P(A \cup B) = P(A) + P(B)$$

5. For events that are *not* mutually exclusive:

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B) \longrightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

6. Venn diagrams can be used to find probabilities

C. Conditional Probability and Independence

1. *Independent Events* – the probability of one event does not change (have an effect on) the probability of another event

2. If A and B are independent then $P(A\ and\ B) = P(A) \cdot P(B) \longrightarrow P(A \cap B) = P(A) \cdot P(B)$

3. To prove that 2 events A and B are independent, show $P(A\ and\ B) = P(A) \cdot P(B)$ *or* $P(B|A) = P(B)$

4. For events that are *not* independent:

$$P(A \text{ and } B) = P(A) \cdot P(B|A) \longrightarrow P(A \cup B) = P(A) \cdot P(B|A)$$

5. Conditional probability formula (use when working with probabilities):

$$P(B|A) = \frac{P(A\ and\ B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

VI. Random Variables

A. Discrete and Continuous Random Variables

1. X = variable whose value is a probability (discrete or continuous)

2. To get the *expected value* or *mean* of a discrete random variable, multiply the number of items by the probability assigned to each item (usually given in a probability distribution table), then sum those products, $\mu = \sum x_i p_i$

3. To get the variance of a discrete random variable, use $\sigma^2 = \sum (x_i - \mu)^2 p_i$ where p is the probability assigned to each item, x.

B. Transforming and Combining Random Variables

1. If $Y = a + bX$ then $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\,\sigma_X$

2. To find the sum or difference ($\pm$) using two random variables, add or subtract the means to get the mean of the sum or difference of the variables, $\mu_{x\pm y} = \mu_x \pm \mu_y$

3. To get the standard deviation ($\pm$) using two random variables, **always add** the *variances* then take the square root of the sum, $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$

C. Binomial and Geometric Variables

1. Conditions for binomial distribution:
   Bi- 2 outcomes (success or failure)
   Nom- Number of observations fixed
   I- Observations independent
   Al- Probability of success is always the same

2. Binomial probability of observing k success in n trials (**binompdf** or **binomcdf**):

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

3. The mean of a binomial distribution is $\mu = np$ where p is the probability and n is the number of observations in the sample.

4. The standard deviation of the binomial distribution is $\sigma = \sqrt{np(1-p)}$

5. The graph of a binomial distribution is strongly right skewed (has a long right tail) unless $n(p) \geq 10 \text{ and } n(1-p) \geq 10$ then the distribution becomes approximately normal.

6. Conditions for geometric distribution are the same as for the binomial except there is not a fixed number of observations because the task is to find out how many times it takes before a success occurs. This is sometimes called a waiting time distribution.

7. The mean of the geometric distribution is $\mu = \frac{1}{p}$

8. The standard deviation of the geometric distribution is $\sigma = \sqrt{\frac{1-p}{p^2}}$

9. The graph of the geometric distribution is strongly right skewed always

10. Geometric probability = $P(Y = k) = (1-p)^{k-1}\, p$ (**geometpdf** or **geometcdf**)