Comprehensive Introduction to Clinical Investigation

Biostatistics - Session Two
Probability and Estimation

Jennifer Gibson, MS
jjgibson@virginia.edu
Division of Biostatistics and Epidemiology
July 5, 2001

Experiments, outcomes, events, and sample space

Experiment

- A process which results in one and only one observation
- A process by which an observation is obtained

Outcome

- The observation resulting from an experiment

Sample space

- Collection of all possible outcomes of an experiment

Example:

Experiment: Rolling 1 die

Outcome: The number on the top of the die

Sample space: All possible numbers on top of the die

Experiments, outcomes, events, and sample space

Event

- Collection of one or more outcomes of an experiment

Simple event (elementary event)

- Includes one and only one outcome

Compound event

- Collection of more than one outcome

Example:

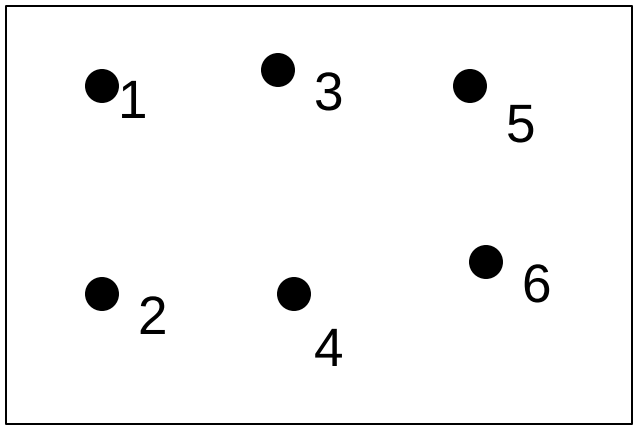
Experiment: Rolling 1 die

Simple event: Observing a 4

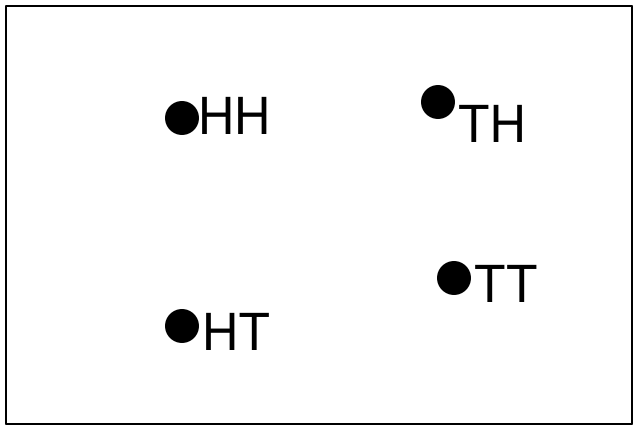
Compound event: Rolling an even number

Venn diagram

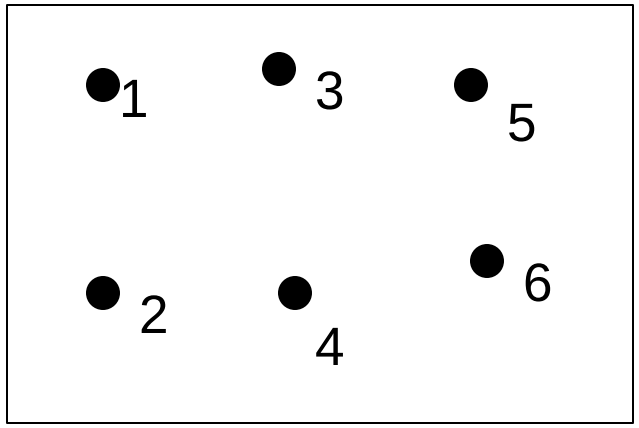
- Picture representation of the sample space of an experiment and events of interest



Roll a die once



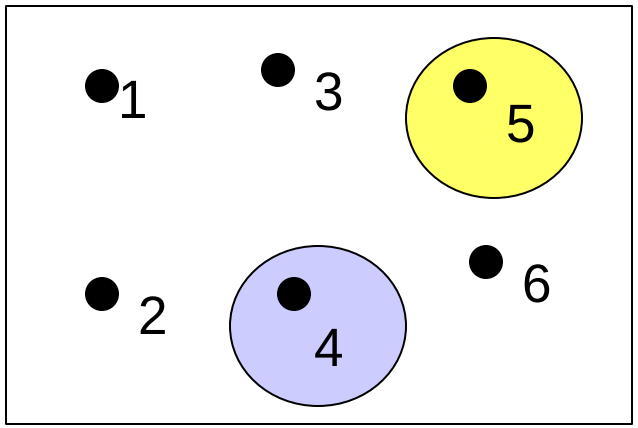
Flip a coin twice



Roll a die once

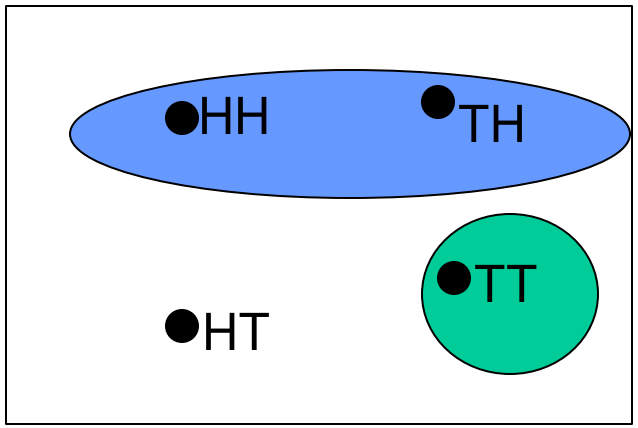
Venn diagram

– Picture representation of the sample space of an experiment and events of interest



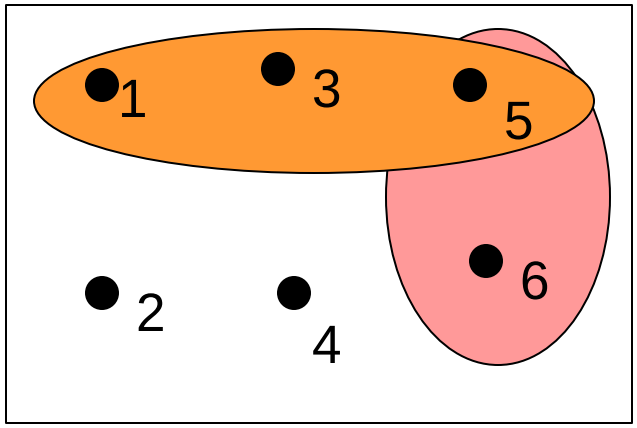
A: Roll a 4

B: Roll a 5



A: Flip 2 tails

B: 2nd flip a head



A: Roll an odd number

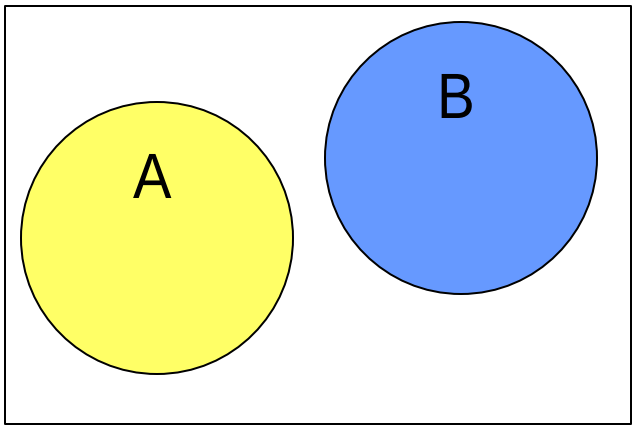
B: Roll a number > 4

Mutually exclusive events

- Events that can NOT occur at the same time

Examples:

- Rolling a 1 on a die and rolling a 6 on the same roll of the die.
- Picking all women for a committee and picking Bob to serve on that committee.

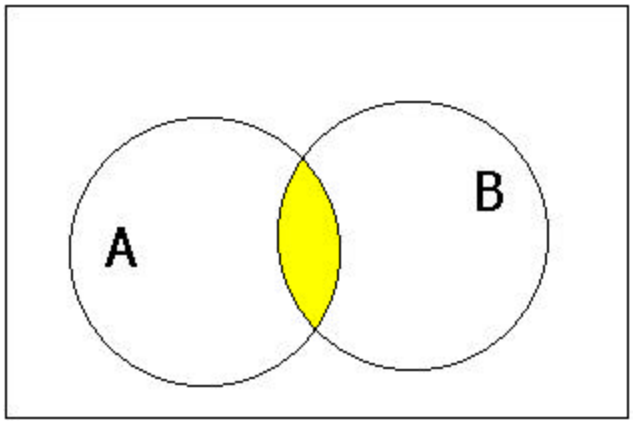


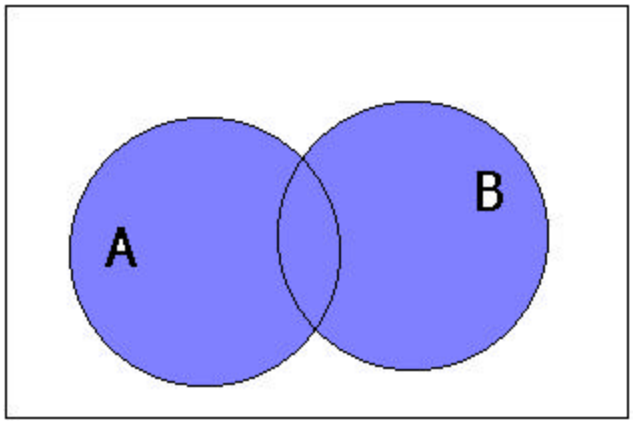
Intersection of events

- The intersection of A and B represents the collection of all outcomes that are common to both A and B
- This is the idea of A AND B both happening

Union of events

- The union of two events A and B includes all outcomes that are in A or in B or in both



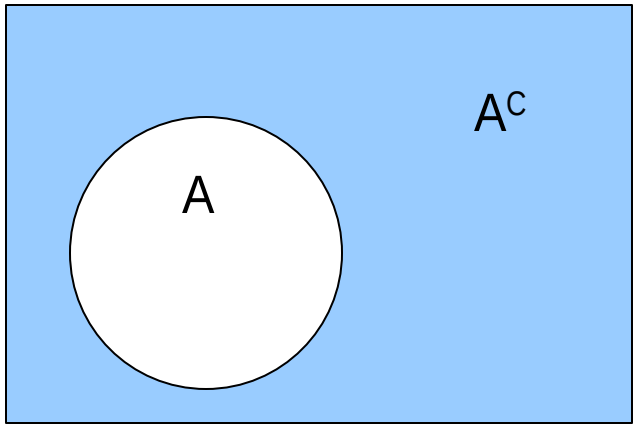


Complementary events

- Two mutually exclusive events that taken together cover the sample space
- Idea of event A and everything except event A

Example:

- A: roll a 2, A^C: roll a 1, 3, 4, 5, or 6



Probability

- Numerical measure of the likelihood of an event occurring
- Allows sample information to be used to make inference to the population
- Notation: P(A) = probability of the event A

Properties

- ALWAYS lies between 0 and 1
- Probability of entire sample space for an experiment is 1
- If P(A)=0, then the event A can NEVER occur during the experiment
- If P(A)=1, then the event A will ALWAYS occur during the experiment

Random variable

- A numerical quantity that takes on different values depending on chance
- Can be discrete or continuous
- Have a distribution consisting of all possible values the random variable can take on and the likelihood of those values occurring
- The likelihood of a value occurring is the probability of the random variable taking on that value

Examples:

- If X is a random variable representing the number of children in a household, $P(X=3)$ is the probability of observing three children in a particular household.
- If Y is a random variable representing the number recorded from rolling a single die, $P(Y=2)$ is the probability of rolling a 2.

Equally likely outcomes

- Two or more outcomes that have the same probability of occurring are equally likely outcomes.
- If all outcomes are equally likely, we can find the probability of an event by counting the number of outcomes that make up the event and dividing by the total number of possible outcomes.

Relative frequency concept

- You could repeat the experiment a large number of times, and then count the number of times the outcome of interest occurs, and use this to estimate the probability of the outcome.
- P(A) = # times A happens / total #experiments
- This is NOT an exact probability but it is an approximation.

Independence

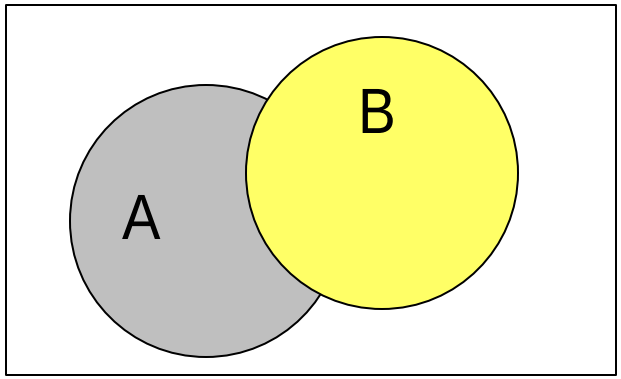
- Two events are independent if the occurrence of one of the events does not change the probability of the other event occurring
- Events can NOT be independent and mutually exclusive
- Multiplication rule: P(A and B) = P(A)*P(B)
- If events are dependent P(A and B) ≠ P(A)*P(B)

Example:

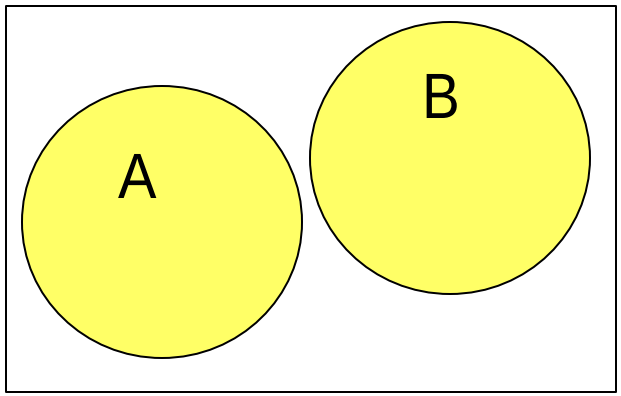
- Experiment: draw 2 balls from a jar with 3 red balls and 2 green balls
  Event A: first ball drawn is red
  Event B: second ball drawn is red
- What if the balls are not replaced between draws?

Probability of the union of two events:

- P(A or B or both) = P(A) + P(B) - P(A and B)



- For mutually exclusive events, P(A and B) = 0, so P(A or B or both) = P(A) + P(B)



Marginal probability

- Probability of a single event without consideration of any other event
- Can use the relative frequency to find marginal probabilities

Conditional probability

- Probability an event will occur given that another event has already occurred. Notation: P(A|B)

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- Once you know that B occurred, what is the likelihood that A also occurred?
- Run into trouble using above formula if P(B)=0

Example:

	Lived	Died	Total
Male	105	26	131
Female	85	14	99
Total	190	40	230

P(male) = # males / total # = 131 / 230

P(lived | female) =

P(lived AND female) / P(female) =

(# lived AND female/total #) / (# female/total #)=

(85/230) / (99/230) = 85 / 99

Example:

	Death penalty		
Gun registration	Favor	Oppose	Total
Favor	784	236	1020
Oppose	311	66	377
Total	1095	302	1397

P(favor gun registration) = 1020 / 1397

P(oppose death penalty| favor gun registration) =
236 / 1020

Independence

- This definition of conditional probability also shows an interesting relationship for independent events.
- Recall, if A and B are independent : P(A and B) = P(A)*P(B)

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) * P(B)}{P(B)} = P(A)$$

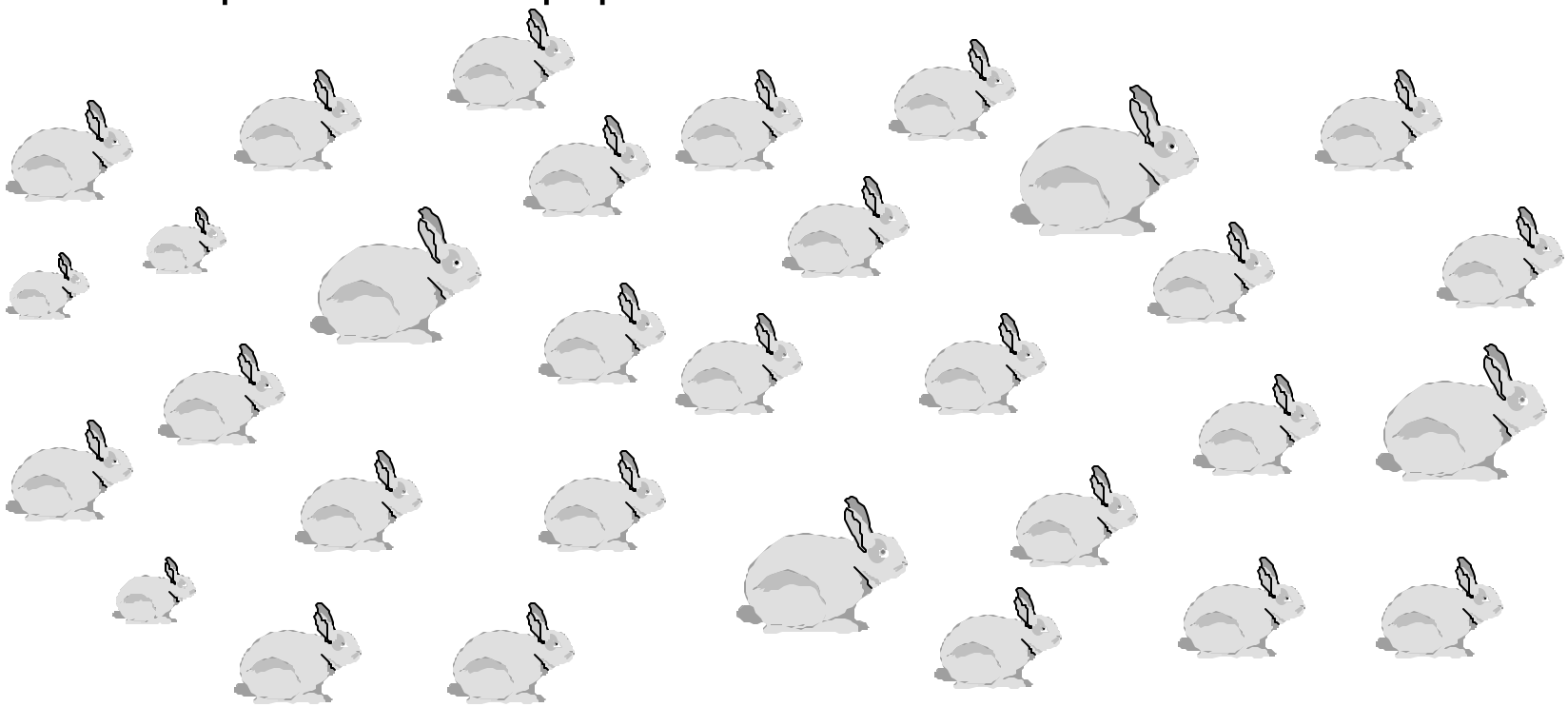
- So we return to the idea that two events are independent when one event (B) occurring does not effect the likelihood of a second event (A) occurring

Population

- Collection of all items of interest

Sample

- A portion of the population on which information is collected

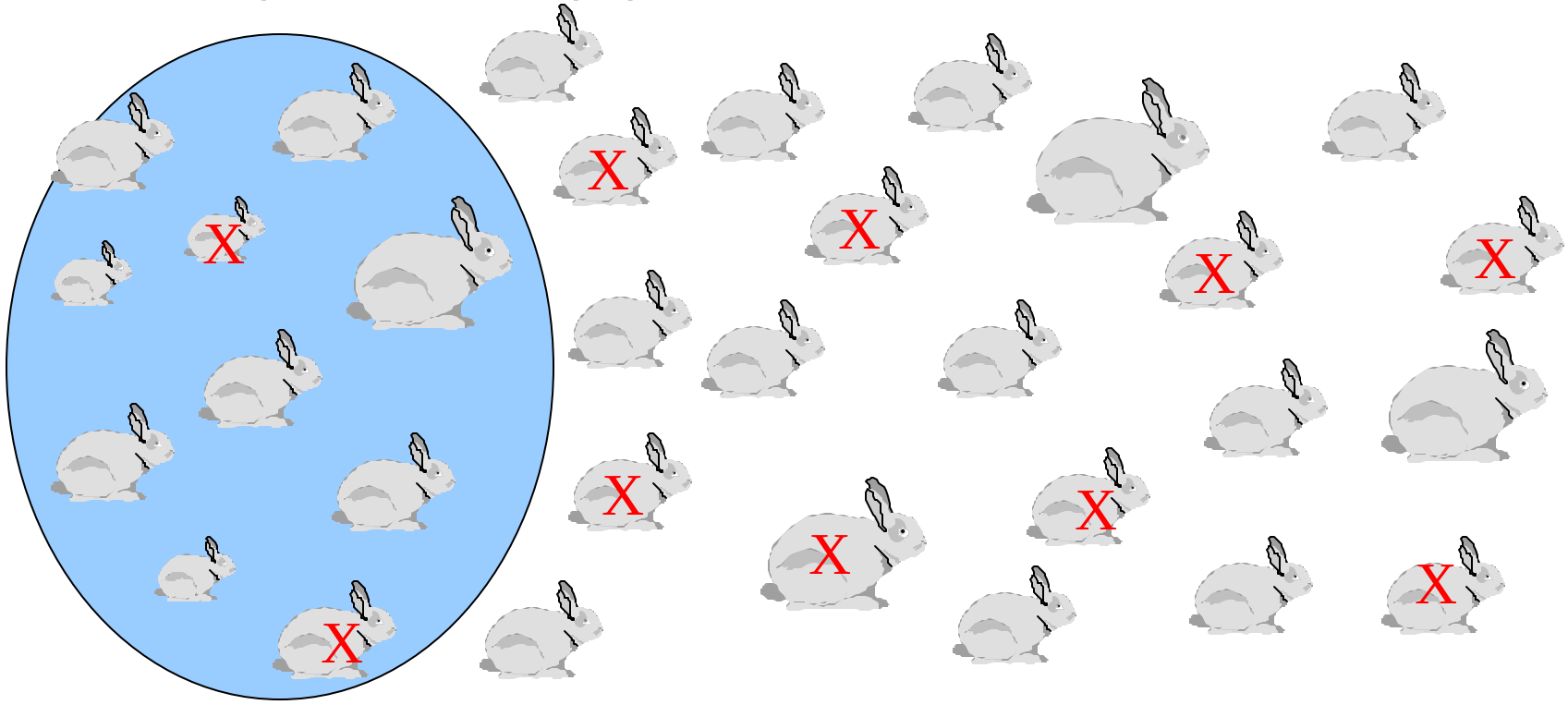


Population

- Collection of all items of interest

Sample

- A portion of the population on which information is collected



Sampling methods

- Random sample

  Every element of the population has **some** chance of being selected. Each element of the sample is chosen independently (without regard to what other elements are chosen).

- Simple random sample

  Every element of the population has the exact **same** chance of being selected.

- With replacement

  Each element **is** returned to the population after sampling. Any particular element can appear repeatedly in the sample.

- Without replacement

  Elements are **not** returned to the population. Any particular element can only appear once in the sample.

Estimation

- The assignment of value(s) to a population parameter based on a value of the corresponding sample statistic
- If we could conduct a census, we would just calculate the population parameter of interest, and we wouldn't need statistics.

Estimate: The value(s) assigned to a population parameter

Estimator: The sample statistic used to estimate the population parameter

- (1) Select a sample and collect information
  (2) Calculate the sample statistic
  (3) Assign value(s) to the population parameter
- Step 3 is called inference. You are inferring the properties of the underlying population based on your sample.

Point estimate

- Results from selecting a single sample and computing a single sample statistic which is used to estimate the population parameter

Interval estimate

- Instead of assigning a single value to the population parameter, you construct an interval and use a probability statement about the presence of the population parameter within the interval

Example:

- Sample of 40 infants born at a hospital. We're interested in the average birthweight of infants at this hospital.
- Point estimate: 117.00
- Interval estimate: [103.4, 111.1]

Quick review of sample statistics (estimators)

Mean (average): $\bar{x} = \frac{\sum x}{n}$

Standard deviation: $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$

Median: value in the sample such that 50% of the sample points are less than or equal to that value (other quantiles calculated similarly)

These estimators are calculated on the sample and then used as point estimates for the same population quantities.

Standard error of the mean

- In order to create interval estimates, we need an idea of how certain we are that the point estimate is correct.

- If we took 100 independent samples from the population, we could calculate the mean for each sample. Then we could calculate a mean of those means and how much each mean differs from the mean of the means. This idea of the difference between each of the 100 means and their mean is the standard error.

- Standard error of the mean is related to the standard deviation of the sample:

$$\text{s.e.m.} = \frac{s}{\sqrt{n}}$$

Interval estimation

- Two factors affect the width of an interval estimate: (1) How big your sample is, (2) how certain you want to be about your interval.

- Since the standard error of the mean is directly related to *n*, the sample size, increasing your sample size will decrease the width of your interval.

- The idea of certainty is related to some statistical ideas that we aren't going to discuss, but in general, the x% confidence interval indicates that you are willing to be incorrect 1-x% of the time. So the more willing you are to be wrong, the narrower your interval. 95% confidence intervals are the most common level, indicating that we are willing to be wrong 5% of the time.

- For a variable that is continuous and has an approximately bell-shaped distribution, confidence intervals are created using the formula: $\bar{x} \pm 1.96 * \text{s.e.m.}$
- The 1.96 is a multiplier (from the normal distribution) that corresponds to the idea of 95% confident. This requires a large (n>30) sample size.

Example:

- Sample of 40 infant birthweights: Mean=117.00, SD=22.44
- s.e.m. = 22.44/$\sqrt{40}$ = 6.95
- Interval: 117.0 $\pm$ 1.96*6.95 = [103.4, 130.6]
- Interpretation: We are 95% confident that this interval covers the population mean.

- New sample of 40 birthweights: Mean=106.70, SD=14.13
- Interval: 106.70 $\pm$ 1.96*2.23= [102.3, 111.1]

Proportions

- For data that are dichotomous, the statistical summary of interest is the proportion.
- $$p = \frac{\sum x}{n}$$
- This is the sample mean for the case of all the observations being 1's or 0's. If we can find the variability of this quantity over a large number of samples, we can use the same formula as for the usual mean.
- The standard error of a proportion is: $\sqrt{\frac{pq}{n}}$
- A 95% confidence interval is given by: $p \pm 1.96 * \sqrt{\frac{pq}{n}}$
- The 1.96 multiplier again depends on a large sample (np>5, nq>5).

Example:

- We would like to estimate the proportion of employees that are female.
- A random sample of 20 health system employees finds 8 females.
- Point estimate of proportion female: p=0.4
- Interval estimate of proportion female:

$$p \pm 1.96 * \sqrt{\frac{pq}{n}}$$

$0.4 \pm 1.96 * \sqrt{(0.4 * 0.6/20)} = [0.19, 0.61]$

- Interpretation: We are 95% confident that the true proportion of females in the health system is between 19% and 61%.

What if p=0?

- If you observe 0 of the events you were looking for, p=0. This is fine for a point estimate.
- But what should we do for an interval estimate? The usual formula we just discussed does not help in this instance, since $\sqrt{(pq/n)} = 0$.

A nice rule of thumb:

- The upper limit of our confidence interval is 3/n
- The lower limit of our confidence interval is 0

- Even though we didn't observe the event yet, it is likely that the event rate is just so small that we didn't observe enough people. The 3/n rule gives us some idea of how big that rate could be, knowing that we didn't see it.

Examples (from Hanly, Lippman-Hand, JAMA 1983):

- 112 live-born children whose mothers had been immunized against rubella were studied for congenital malformations. None of the infants showed malformation.

  An interval estimate for the proportion of infants with malformations is [0, 3/n] = [0, 3/112] = [0, 0.027]

  Interpretation: We are 95% confident that the malformation rate in this group does not exceed 2.7%

- 14 boys after chemotherapy for leukemia were studied for abnormal testicular function. None of the boys showed abnormal function.

  An interval estimate for the proportion with abnormal function is: [0, 3/14] = [0, 0.21]

  Interpretation: We are 95% confident that the risk of abnormal testicular function does not exceed 21%.

Additional examples

Two cold tablets are accidentally placed in a box containing two aspirin tablets. The four tablets are identical in appearance. One tablet is selected at random and taken by patient A. One of the remaining tablets is select at random and taken by patient B.

P(A: B took a cold tablet)

P(B: exactly one of the patients took a cold tablet)

P(C: neither patient took a cold tablet)

P(A: and B:)

Suppose we wish to estimate the concentration of a specific dose of ampicillin in the urine after a certain period of time. We recruit 35 volunteers and measure the concentration of ampicillin in their urine.

- Mean concentration=7.0 μg/mL
- SD concentration=2.0 μg/mL

- Point estimate of population mean=7.0 μg/mL
- 95% confidence interval for population mean:

  $7.0 \pm 1.96*(2.0/\sqrt{35})$

  [6.34, 7.66]
- Interpretation: We are 95% confident that the true population mean concentration of ampicillin in the urine of patients lies between 6.34 and 7.66 μg/mL.

Additional examples

Suppose a clinical trial is conducted to test the efficacy of a new drug in the treatment of gonorrhea for females. Forty six patients are given a 4g daily dose of the drug and one week later observed for evidence of gonorrhea.

- Number with gonorrhea at 1 week = 6
- Proportion with gonorrhea = 6/46 = 0.13

- Point estimate of drug failure rate = 0.13
- 95% confidence interval for population failure rate:

  $0.13 \pm 1.96*\sqrt{(0.13*0.87/46)}$

  [0.03, 0.23]
- Interpretation: We are 95% confident that the population failure rate lies between 3% and 23%.