

LECTURE NOTES

BIOE 811: BIostatISTICS FOR THE HEALTH SCIENCES I

COURSE DIRECTOR: BETSY TOLLEY, PH.D.

BIOE 811
BIOE 840, Section 001

BIostatistics FOR THE HEALTH SCIENCES I
Special Topic—Biostatistics Laboratory
FALL, 2009

Professor: Betsy Tolley, Ph.D.
Office: Room 613, 66 N. Pauline Street, Doctors Office Building
E-mail: btolley@utmem.edu
Phone: 448-5137 or 448-5900 (for receptionist)
Office Hours: After class or by appointment
Class Hours: 8:15 - 9:15 A.M. Monday, Tuesday, and Friday.
Laboratory Hours: 9:30 – 10:30 A.M. Monday
Class Location: Room 601, 877 Madison Avenue, Alexander Building

Texts: Bernard Rosner, Fundamentals of Biostatistics, 6th Ed.

Recommended Supplementary Texts:

Robert R. Sokal and F. James Rohlf. Biometry: The Principles and Practices of Statistics in Biological Research, 3rd Ed.

Lora D. Delwiche and Susan J Slaughter, The Little SAS[®] Book: A Primer, 3rd Ed or 4th Ed.

Sandra D. Schlotzhauer and Ramon C. Littell. SAS[®] System for Elementary Statistical Analysis, 2nd Ed.

Lecture Notes: Lecture notes are available on the Preventive Medicine website <<http://www.utmem.edu/prevmed/>>. You will need to have a two-inch three-ring binder in which to hold your lecture notes. You may either print the pdf files to three-hole punched paper or hand-punch regular paper to fit in your binder. Dividers may be useful in separating various portions of your lecture notes. Hand-written notes that you take in class may be easily placed with your printed lecture notes.

Homework: Problems will be assigned to give you experience in obtaining certain statistics, interpreting descriptions of studies, or using SAS. These assignments are contained in the lecture notes. Unless otherwise specified, all assignments must be turned in for credit. Homework problems should be submitted to Dr. Tolley on the Friday following the class discussion of that assignment (usually done in the laboratory setting on Mondays). Full credit will be given for complete assignments only; incomplete assignments will receive half credit. Furthermore, only half credit will be given for any homework assignment turned in more than 2 weeks after the corresponding class discussion. Prior arrangements must be made if a deadline cannot be met. Note that homework assignments will not be checked for accuracy. Each student is responsible for making appropriate corrections to his or her work during class discussions. The homework will provide the basis for your grade (P/NP) in the laboratory portion of this course.

Exams & Grading: There will be three exams, all of which will be both open-book and open-note. These exams will be administered during the class period (i.e., **not** “take-home”). Final grade is determined by module 1 (exam 1) (25%), module 2 (exam 2) (25%), module 3 (exam 3) (25%), and homework (25%).

Calculator: You will need to have a calculator that has a Σ -key and performs advanced statistical functions.

Reading: Assignments should be read before scheduled class discussions.

Computer: You must have access to a desktop or laptop personal computer so that you can complete your assignments using Microsoft Excel[®] and SAS[®] statistical software. There are a limited number of desktop personal computers available for student use in the Preventive Medicine Computer Laboratory in Room 600.

Excel[®]: Throughout this course, you will need to have access to Microsoft Excel[®] on a desktop or laptop computer.

SAS[®] Software: You must purchase or have access to SAS[®] software. A license must be purchased for \$48.07 or \$59.00 from UTHSC General Stores. Directions for obtaining this license have been distributed separately.

MODULE 1

A. General Overview and Descriptive Statistics

1. Ch. 1. General Overview, pp. 1-5.
Ch. 2. Descriptive Statistics, Sections 2.1-2.6, pp. 6-24.
 - a. Measures of Central Location
 - b. Properties of the Arithmetic Mean
 - c. Measures of Spread
 - d. Properties of the Variance and Standard Deviation
 - e. The Coefficient of Variation

2. Ch. 2. Descriptive Statistics, Sections 2.7-2.9, pp. 25-33.
 - f. Grouped Data
 - g. Graphic Methods for Displaying Data

B. Probability

3. Ch. 3. Probability, Sections 3.1-3.6, pp. 43-57.
 - a. Probability: Definition and Notation
 - b. Independent and Dependent Events
 - c. Multiplication and Addition Laws of Probability
 - d. Conditional Probability

C. Probability Distributions

4. Ch. 5. Continuous Probability Distributions, Sections 5.1-5.5, pp. 122-139.
 - a. The Normal Distribution

D. Estimation

5. Ch. 6. Estimation, Sections 6.1-6.5, pp. 166-194.
 - a. Relationship between a Population and a Sample
 - b. Random Number Table
 - c. Estimation of the Mean of a Distribution

6. Ch. 6. Estimation, Sections 6.6-6.7, pp. 195-201.
 - d. Estimation of the Variance of a Distribution

E. Hypothesis Testing

7. Ch. 7. Hypothesis Testing, Sections 7.1-7.4, pp. 226-245.
 - a. One-sample Test for the Mean of a Normal Distribution – One- and Two-sided Alternatives

8. Ch. 7. Hypothesis Testing, Section 7.5, pp. 245-252.
 - b. The Power of the Test

9. Ch. 7. Hypothesis Testing, Sections 7.6-7.7, pp. 253-262.
 - c. Sample Size Determination
 - d. The Relationship between Hypothesis Testing and Confidence Intervals

10. Ch. 7. Hypothesis Testing, Sections 7.9-7.11, pp. 267-270.
 - e. One-Sample Chi-Square Test for the Variance

**** Module 1 Exam 1 – Fundamentals of Biostatistics – Chs. 1, 2, 3, 5, 6, and 7.**

MODULE 1 – LECTURE NOTES

INTRODUCTION

The **Scientific Method** requires researchers to ask and investigate research questions, formulate or specify hypotheses, choose those hypotheses that their data support, and reject the others.

Statistical methods allow researchers (e.g., basic scientists and clinical investigators) to compute statistics according to a set of rules based on the Scientific Method.

DESCRIPTIVE STATISTICS

Definitions:

Statistics – the science of collecting, analyzing interpreting and presenting data; involves data reduction.

Biostatistics – the application of statistical methods to biological or biomedical problems.

Data – numerical observations or measurements that are results of random phenomenon.

Random Phenomenon – the investigator cannot reproduce any data exactly; always some natural variation.

Random (stochastic) – something that is not controlled; a certain amount of variability always present.

Natural variation means that data are irreproducible. A situation in which there is always variability causes **uncertainty**. In order to identify reasons for observed differences, the researcher must sort out the **special causes** that lead to **systematic variation**, separating those from the natural variation that is always present. Decisions will be uncertain.

Statistical inference – an objective way of evaluating data to make decisions when there is uncertainty.

Statistical models – aid the researcher in making **inferences** and decisions based on data; used to represent data in terms of special causes and natural variation.

Contrast the stochastic situation with the **deterministic** situation.

Deterministic – no natural variability.

Deterministic data are reproducible.

Deterministic decisions are certain.

Mathematical models, such as differential or integral equations, represent data exactly.

Uses of Statistics:

Information	Variability	
	None	Present
Full (Population)	None	Descriptive
Partial (Sample)	None	Inferential

Random variables are recorded in the form of numbers.

Certain statistical techniques apply to each type of random variable.

a) **Measurement variables**

- 1) **Continuous**
- 2) **Discrete**

b) **Ranked or ordinal variables**

c) **Attributes** must translate into numbers, e.g., frequencies of occurrence

d) **Computed variables**

Two important issues relevant to the use of statistics are **accuracy** and **precision**.

Accuracy

- closeness of the measure to the true value;
- has to do with bias.

Precision

- closeness of repeated measurements;
- has no bearing on closeness to the true value.

Precision without accuracy can be a problem when we are trying to make statistical inferences.

DESCRIPTIVE MEASURES

Measures of Location

Location refers to where on an axis a particular group of data is located relative to a norm or another group.

- 1) **Mean** – arithmetic mean, expected value, average; has the most use, but is sensitive to extreme values.

Other means – **geometric mean**,
harmonic mean.

- 2) **Median** – 50% point, 50th percentile, the middle observation; insensitive to extreme values, but determined primarily by middle values.

- 3) **Mode** – the value with the highest frequency; a measure of concentration.

Population Parameters: Measures of Location

$Y_i = Y_1, Y_2, \dots, Y_N$ are the individual observations in a population.

$$\begin{aligned}\mu &= 1/N (Y_1 + Y_2 + \dots + Y_N) \\ &= 1/N \sum_{i=1}^N Y_i\end{aligned}$$

Measures of Dispersion

- 1) **Range** – distance between the highest (largest) and lowest (smallest) value;
 - a) **Interquartile range** – distance between the 25% and 75% points.

Range corresponds to the median.

Symbolized by “R”.

- 2) **Standard deviation** – the average distance from the mean; if the standard deviation is small, the observations are crowded near the mean; if the standard deviation is large, there is substantial spread in the data.

Standard deviations correspond to means.

One function of the standard deviation is to give the variance.

Symbolized by “ σ ”.

3) **Variance**

$$\sigma^2 = (\text{standard deviation})^2 \\ = \text{mean square.}$$

The variance (rather than the standard deviation) is used in calculations.

$$\sigma^2 = 1/N [(Y_1 - \mu)^2 + (Y_2 - \mu)^2 + \dots + (Y_N - \mu)^2] \\ = 1/N \sum_{i=1}^N (Y_i - \mu)^2$$

Ignore an expression in the texts called the computing formula.

Other Descriptive Measures

1) Measure of **Skewness**

2) Measure of **Kurtosis**

Coefficient of Variation

Population: $\sigma/\mu * 100\% = CV$

Makes σ in various populations comparable because these have been standardized by the mean.

Gives an indication of how reliable and reproducible an experiment is.

Because we want the σ or spread as small as possible relative to the mean, it may be necessary to improve the measurement technique.

Parameters versus Statistics

Greek letters are used to denote **parameters** of a **population**.

These parameters are generally unknown.

Parameters are not statistics.

Usually investigators do not have the entire population.

Instead they have partial information and want to “infer” something about the population.

A **sample** is a subset of the population.

A **random sample** gives **estimates**, which are **statistics**, for the population parameters.

Thus, statistics are estimates for corresponding parameters from a **reference population**.

In a sample the observations are now:

Y_1, Y_2, \dots, Y_n , where $n < N$.

Sample mean:

$$\begin{aligned}\bar{Y} &= 1/n (Y_1 + Y_2 + \dots + Y_n) \\ &= 1/n \sum_{i=1}^n Y_i, \text{ where } n = \text{sample size.}\end{aligned}$$

This estimate of the population mean will give the best information about μ .

This statement implies that the sample mean is an **unbiased estimate** of the population mean.

Sample variance:

$$\begin{aligned}s^2 &= 1/(n-1)[(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \dots + (Y_n - \bar{Y})^2] \\ &= 1/(n-1) \sum_{i=1}^n (Y_i - \bar{Y})^2\end{aligned}$$

The sample variance is an unbiased estimate of σ^2 .

Ignore an expression in the texts called the computing formula.

$$s = \sqrt{s^2}$$

$(n-1)$ is called degrees of freedom (d.f.).

Sample coefficient of variation:

The variance s^2 is always positive.

\bar{Y} does not have to be non-negative.

CV for sample = s/\bar{Y} .

Generally, we take the CV as the absolute value of s/\bar{Y} , i.e., even if \bar{Y} is negative, the CV is positive.

The CV is used to make s in several samples comparable.

A property of biological data is that for a given variable the s usually increases (decreases) with \bar{Y} .

The result is that for a particular variable the CV tends to be reasonably constant over a wide range of values.

FREQUENCY DISTRIBUTIONS FOR SAMPLES

Interval	Limits	Class mark (Midpoint)	Abs. freq.	Rel. freq.	Cum. freq.
I_1	$a_0 - a_1$	$Y_1 = (a_0 + a_1)/2$	f_1	$P_1 = f_1/n$	$C_1 = P_1$
I_2	$a_1 - a_2$	Y_2	f_2	P_2	$C_2 = P_1 + P_2$
I_3	$a_2 - a_3$	Y_3	f_3	P_3	$C_3 = P_1 + P_2 + P_3$
I_t	$a_{t-1} - a_t$	Y_t	f_t	P_t	$C_t = P_1 + \dots + P_t = 1.0$
			n	1.0	

Rule of Thumb: $t = \text{approx. } \sqrt{n}$, where n = the total number of observations.

Steps:

- 1) Determine the number of **intervals** and calculate the **limits**, a_i .
 Generally, limits are the same size.
 General rule for interval length or size: (maximum value – minimum value)/ t .
 If there are too few intervals, then the grouped data are not reflective of the nature of the data.
 If there are too many intervals, there has been no data reduction.
- 2) Calculate the **class marks**, Y_i .
 These are equal to the midpoint of the interval and are also called the **value of the interval**.
- 3) Partition observations into different intervals.
 Count the number of observations in each interval and record the **absolute frequencies**, f_i .
 Absolute frequencies are the number of observations in each interval.
 An observation can be assigned to one and only one interval.
 If an observation is equal to the limit, arbitrarily assign it to either the lower or upper interval.
 Be consistent within each data set so that all border cases are either assigned up or down.
- 4) Calculate **relative frequencies**, P_i .
 Relative frequencies are the proportion of observations in each group or interval.
 Relative frequency is one of the most important concepts in statistics.
 The total of the relative frequencies must equal 1.0 (100%).
- 5) Calculate **cumulative frequencies**, C_i , by adding the relative frequencies from class one to the next. Then, add this sum to the next relative frequency. Continue this procedure.
 The final cumulative frequency will be equal to 1.0 (100%).

- 6) Multiply each absolute frequency times the respective class mark, $f_i Y_i$.
 Add these up, $\sum_{i=1}^t f_i Y_i$, and divide by the sum of the frequencies ($\sum_{i=1}^t f_i = n$).
 The result is equal to the **grouped mean**, \bar{Y}_g .
- 7) Calculate the **grouped standard deviation** by using one of the formulas in the text or the formula depicted below.
 As an alternative method, these calculations can be done simply by adding three more columns to the chart: $(Y_i - \bar{Y}_g)$, $(Y_i - \bar{Y}_g)^2$, and $f_i(Y_i - \bar{Y}_g)^2$.
Remember: In this case and most other practical situations, the standard deviation is calculated from a sample so the denominator is $n-1$.

Grouped Mean:

$$\bar{Y}_g = \frac{\sum_{i=1}^t f_i Y_i}{\sum_{i=1}^t f_i}$$

where f_i is the frequency in the i -th interval and Y_i is the class mark of the i -th interval.

Grouped Variance:

$$s_g^2 = \frac{\sum_{i=1}^t f_i (Y_i - \bar{Y}_g)^2}{[\sum_{i=1}^t f_i - 1]}$$

where f_i , Y_i , and \bar{Y}_g are defined as the frequency in the i -th interval, the class mark of the i -th interval, and the grouped mean, respectively.

- 8) Although not always required, a **graph of the data** can be very helpful, especially during **exploratory** or **preliminary data analysis**.

Graphical Depiction of Data from Small Samples

Histograms and **bar graphs** are often used to picture the data.

The **shape of the graph** will be the same regardless of whether absolute or relative frequencies are used.

Shape can be important; it gives some information.

Cumulative frequencies can also be used.

In this case the step size is important.

Graphs of cumulative frequencies can be a little more difficult for the inexperienced researcher to decipher than graphs of absolute or relative frequencies.

Often graphs of cumulative frequencies are used expressly for the purposes of obscuring large differences in frequencies at the lower or upper end of a scale.

Graphical Depictions of Data from Large Samples or Populations

For large samples or populations with infinitesimally small class sizes, the histogram can be represented by a **smooth curve**.

This is now called the **distribution** of the sample or population and has all properties of the histogram.

These curves can be represented by mathematical equations or functions, which can be used to get proportions.

The **normal distribution** is one of the most important of these curves.

It is an idealized representation of the “true” situation.

Data will not represent a normal curve or distribution exactly.

Many of the statistics and statistical models that we use depend on the **assumption** that data come from a normal distribution.

If the fit is so poor that data do not even approximate a normal distribution, we must use other methods for analyzing those data.

PROBABILITY

An **event** refers to any particular outcome or set of outcomes that we are interested in studying.

The **probability** of an event occurring, p , is the relative frequency for this outcome or set of outcomes, if an infinite number of trials were made.

The sum of the probabilities for all possible outcomes is equal to 1.0.

The probability that a particular event, $\{E\}$ will occur – $\Pr(E)$ – is greater than or equal to zero and less than or equal to 1.0

$$0 \leq \Pr(E) \leq 1.0$$

If two events $\{A\}$ and $\{A^c\}$ are **mutually exclusive**, $\Pr(A \text{ or } A^c) = \Pr(A) + \Pr(A^c)$.

Disease:

<u>Present</u>	<u>Absent</u>
p	1-p

Total: $p + (1 - p) = 1.0$

The probability of an event occurring is a quantitative expression of the **likelihood** of its occurrence.

Probability is best defined in terms of relative frequency:

$$\Pr(A) = p = \frac{\text{\# of times A does occur}}{\text{total \# of times A can occur}}$$

Probabilities can be expressed as fractions, decimal fractions, or percentages.

When expressed as decimal fractions, probabilities fall in the range of 0 to 1.0.

In biomedical studies we are often interested in looking at more than one event at a time.

For example, we may be interested in studying families, defining $\{A\}$ as parents and $\{B\}$ as children.

Now we are concerned with $\{A\}$ and $\{A^c\}$ as well as $\{B\}$ and $\{B^c\}$.

Definition of Intersection:

$\{A \cap B\}$ is defined as the event that both A and B occur simultaneously or jointly.

Test for Independence Using the Intersection Rule:

Two events are independent, i.e., occur together by chance alone, if

$$\Pr(A \text{ and } B) = \Pr(A \cap B) = \Pr(A) * \Pr(B).$$

If the probability of the two events occurring together (i.e., the **joint probability**) is not equal to the product of their separate probabilities, then the events are dependent.

Definition of Union:

$\{A \cup B\}$ is defined as the event that either A or B occurs or they both occur.

There are two situations:

1. A and B are mutually exclusive.

2. A and B are not mutually exclusive.

Union or Addition Rule:

For any two events the probability of A occurring, or B occurring, or both occurring is
 $\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Test for Independence Using the Union Rule:

If A and B are independent, then the probability of A, or B, or A and B occurring is
 $\Pr(A \text{ or } B | \text{Independence}) = \Pr(A) + \Pr(B) * [1 - \Pr(A)]$

$$= \Pr(A) + \Pr(B) * \Pr(\bar{A}).$$

If the probability of A occurring, or B occurring, or both occurring is not equal to this quantity, then the events are dependent.

Definition of Conditional Probability:

A conditional probability is the probability that B occurs given that A has occurred.

$$\Pr(B|A) = \frac{\text{\# of times A and B occur jointly}}{\text{\# of times A occurs}} = \Pr(A \cap B) / \Pr(A).$$

Thus, by rearranging terms, the joint probability can be expressed as a product of conditional probability and a total probability:

$$\Pr(A \cap B) = \Pr(B|A) * \Pr(A).$$

Test for Independence Using the Conditional Probability:

If events A and B are independent, the occurrence of A does not influence the occurrence of B.

Therefore, $\Pr(B|A) = \Pr(B)$.

Addition or Total Probability Rule:

For any two events {A} and {B}, the **total** or **marginal probability** of {A}, i.e., $\Pr(A)$, equals the sum the joint probability of A and B and the joint probability of A and \bar{B} :

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \bar{B}).$$

Thus, for any situation, by substituting the product of a conditional probability and a total probability for the respective joint probability,

$$\Pr(A) = \Pr(A|B)*\Pr(B) + \Pr(A|\bar{B})*\Pr(\bar{B}).$$

Test for Independence Using the Total Probability Rule:

For independence,

$$\Pr(A) = \Pr(A) * \Pr(B) + \Pr(A) * \Pr(\bar{B}).$$

If the total probability of A is not equal to this quantity, then the events are dependent.

CONTINUOUS PROBABILITY DISTRIBUTIONS

A **continuous random variable**, X , is any random variable that is not discrete.

The values of a continuous random variable form a continuum.

The probability of exactly obtaining any particular value is zero.

The **probability density function** of a random variable, X , is a curve such that the area under the curve between any two points, a and b , is equal to the probability that X falls between a and b .

The total area under the curve over the possible range of the random variable, which theoretically extends from $-\infty$ to $+\infty$, is 1.0.

The cumulative distribution function for X at the point a is defined as the probability that X will have values $\leq a$: $\Pr(X \leq a)$.

Graphically, the cumulative distribution function is depicted by the area under the probability density function to the left of a .

Most of the estimation procedures and hypothesis testing techniques commonly used in biostatistics depend on the assumption that the random variable that we are studying has an **underlying normal distribution**.

Many random variables, which are not normally distributed, may be transformed so that the transformed variable is.

Other random variables, e.g., growth rate, are affected by many factors. We can decompose such a random variable into the sum of several other random variables. Often, this sum is normally distributed.

In addition, we make some random variables conform to the normal distribution by describing **special causes** and having the part of the variable that remains unexplained follow the normal distribution.

This means that we can use the normal distribution because any statistics obtained from samples are random variables.

These statistics are estimates of real parameters and have some variability.

Usually we want to say something about the corresponding parameters.

We should remember that the normal distribution is a mathematical concept used to approximate what is expected in nature.

But, it is flexible enough to be representative of a very wide range of biological and biomedical data, which have the same structure.

Normal Distribution

$$Y \sim N(\mu, \sigma^2)$$

Random variable, Y , is distributed normally with a mean μ and a variance σ^2 .

μ and σ^2 are unknown and unknowable.

The curve is completely determined by μ and σ^2 .

$$E(Y) = \mu \text{ and } \text{Var}(Y) = \sigma^2.$$

The **normal probability density function** or **normal distribution** is a bell-shaped curve.

The mean, mode and median are all equal for the normal distribution.

This curve is the limiting form of a histogram with an infinite number of classes.

It is a reasonable approximation to the actual frequency distribution we would have with a very large number of classes.

The normal curve is symmetrical about μ and the points of inflection are located at $\mu \pm \sigma$.

Because there are an infinite number of μ 's and σ^2 's, there are an infinite number of curves.

Usually we need to know the proportion of the population $<$ or $>$ certain values.

We could obtain this value by integration or by using tables for all these curves.

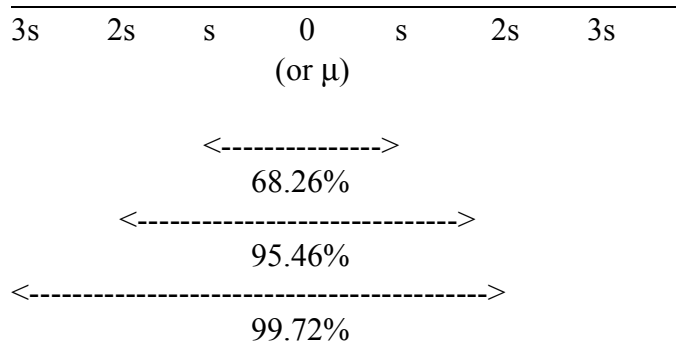
However, to use a single table, we need to define the **standard normal distribution**.

A **standardized random variable** is defined as $X = (Y - \mu)/\sigma$ where $Y = \mu + X\sigma$.

A standardized random variable is sometimes called “z” or “z-scores”.

$X \sim N(0, 1)$ so that $\mu_X = 0$ and $\sigma_X^2 = 1$.

Once we know something about X, we can go back to Y.



Using these properties we can determine if an individual measurement came from one population or perhaps another.

The research question usually involves determining this.

The population is loosely defined by a set of conditions, e.g., normal or abnormal.

The **cumulative distribution function for the standard normal distribution** is denoted $\phi(x) = \Pr(X \leq x)$, where $X \sim N(0, 1)$.

From the symmetry properties of the standard normal curve,
 $\phi(-x) = \Pr(X \leq -x) = \Pr(X \geq x) = 1 - \Pr(X \leq x) = 1 - \phi(x)$.

In many cases we are interested in the tail regions on either side of 0 for the standard normal distribution.

Conceptually, the body and tail regions correspond to that part within and outside the “normal” or “expected” ranges of the population in which we are interested.

The probability of a value falling in this range is given by $\Pr(-x \leq X \leq x)$.

The 100% * p-th percentile of a standard normal distribution, denoted by z_p , is defined as $\Pr(X < z_p) = p$ where $X \sim N(0,1)$.

Uses of the Standard Normal Table

Example:

$Y \sim N(15, 36)$

1. What proportion of the population is larger than 22?

$\mu = 15; \sigma^2 = 36; \sigma = 6.$

Between 9 and 21 lies approximately 60%.

$\Pr[Y \geq 22] = \Pr[(Y - 15)/6 \geq (22 - 15)/6] = \Pr[X \geq 1.17]$, where $X \sim N(0, 1)$.

$(Y - 15)/6$ is an example of a standardized normal random variable.

$(22 - 15)/6$ is the value of that standardized normal random variable or the z-score.

Look at table 3:

From column A, the area to the left of 1.17 is .8790 or 87.9% of the population.

From column B, the area to the right of 1.17 is .121 or 12.1% of the population.

From column C, the area between 0 and 1.17 is .379 or 37.9% of the population.

From column D, the area between -1.17 and 1.17 is .758 or 75.8% of the population.

The answer comes from column B: $\Pr[X \geq 1.17] = \Pr[Y \geq 22] = .121$.

Thus, we can expect 12.1% of the population to have values ≥ 22 .

2. What is the value k^* such that 98% of the population lies below k^* , i.e., is less than k^* ?

Now, k^* is the value on our original scale; k is our value on the z-score scale.

$$Y \sim N(15, 36)$$

$$\Pr[Y \leq k^*] = .98$$

$$\Pr[(Y - 15)/6 \leq (k^* - 15)/6] = .98$$

$$\Pr[X \leq k] = .98$$

Look at table 3.

Find .98 under column A.

This value lies between .9798 and .9803 and corresponds to the interpolated x-value of 2.054.

$$k = 2.054$$

$$k = (k^* - 15)/6$$

$$k^* = 15 + k(6) = 15 + 2.054(6) = 27.324$$

STATISTICAL INFERENCE

When we want to infer something about the properties (i.e., parameters) of an underlying distribution of a population based on the properties (i.e., estimates of parameters or statistics) of a sample that we have studied, we use **inductive logic and reasoning**.

This process, called **statistical inference**, has two main subdivisions:

1. **Estimation** is concerned with estimating the value of a specific population parameter.
2. **Hypothesis testing** is concerned with testing whether the value of a population parameter is equal to, less than, or greater than some specified value or **norm**.

The relationship between a sample and the population affects the **external validity** of the findings.

External validity is the ability to validly apply the findings based on the sample that we have studied to the population that we wish to study.

How the sample is obtained affects external validity.

A **random sample** is obtained by selecting some members of the population in a manner such that each member is independently chosen and has a known non-zero probability of being selected.

A **simple random sample** is a random sample in which each member has the same probability of being selected.

In the scientific literature, the term *random sample* usually implies *simple random sample*.

The **reference, target, or study population** is the group that we wish to study. In theory, the random sample is selected from this group.

Random samples are not the only samples used in health science research.

Other types of sampling include cluster sampling (described in the text) and systematic sampling (e.g., every other item, every tenth item).

In practice, the opportunity rarely exists to enumerate every member of the population for the purpose of selecting a random sample.

Therefore, we make the **assumptions** that (1) the sample selected has all the properties of a random sample without formally being a random sample and (2) the reference population is effectively infinite.

Whether these are good or bad assumptions affects external validity.

RANDOM-NUMBER TABLES AND GENERATORS

A **random number** is a random variable, X , that takes on the values of 0, 1, 2, ..., with equal probability.

A **random-number table** is a collection of digits that satisfies the following two properties:

1. Each digit 0, 1, ..., 9 is equally likely to occur; and
2. The value of any particular digit is independent of the value of any other digit in the table.

Many **random number generators** exist; various types of random number generators have been used throughout history in games of chance. Any of these generators can be used to generate a table of random numbers. By using a **pseudo-random number generator**, computer programs generate large sequences of random digits that approximately satisfy the required conditions.

Random-number tables may be used for **random selection** of subjects, animals, or items or for **random assignment** of treatments to subjects, animals, or items.

Randomization implies random assignment of treatments to subjects, animals, or items.

Block randomization is often used to restrict randomization to blocks of size s in order to ensure that the treatments have equal numbers of subjects, animals, or items from the beginning of the study.

For example, two treatments randomly assigned in blocks of four results in 6 ways in which the subjects, animals, or items can be randomized (i.e., ${}_4C_2$):
C C T T, C T C T, C T T C, T C C T, T C T C, and T T C C.

Stratification is the process of subdividing subjects, animals, or items according to characteristics that are associated with the response variable.

After stratification, random selection or random assignment is conducted for each stratum.

DISTRIBUTION OF SAMPLE MEANS

Suppose we have r samples of size n , where n = number of observations selected from an infinitely large population:

$$\begin{array}{l} Y_{11}, Y_{12}, \dots, Y_{1n}, \\ Y_{21}, Y_{22}, \dots, Y_{2n}, \\ \\ Y_{r1}, Y_{r2}, \dots, Y_{rn}, \end{array}$$

For each sample we can calculate a sample mean, \bar{Y}_r .

The sampling distribution of \bar{Y}_r is the distribution of values of \bar{Y}_r over all possible samples of size n that could have been selected from the original population.

Now, we can have an infinite number of observations in our new population of means.

We can obtain a **mean of sample means**: $\mu_{\bar{Y}} = \mu =$ mean of original population.

Note that the mean of sample means is the same as the mean of the original population of individual values.

Thus, based on the one sample we actually make, \bar{Y} is the point estimate of μ : $\hat{\mu} = \bar{Y}$ or, alternatively, $E(\bar{Y}) = \mu$.

The **variance of sample means** is needed to give an indication of dispersion or spread among all possible means from all possible samples: $\sigma_{\bar{Y}}^2 = \sigma^2/n$.

Thus, the variance of sample means is $1/n$ * variance of the original observations.

The **standard error** is the standard deviation associated with the population of means (the standard deviation of the mean): $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$

If n gets very large, the standard error gets very small and will approach zero.

To obtain a very large sample, we will have sampled the entire population each time and all the samples will be the same. Therefore, \bar{Y} is no longer a random variable, but a constant and no variance exists.

What about estimates of dispersion from one sample that we actually make?

Based on the one sample we actually make, $s_{\bar{Y}} = s/\sqrt{n} =$ the **estimated standard error**, usually called “the standard error”.

S.E.(\bar{Y}) = \sigma_{\bar{Y}} = \sigma/\sqrt{n} is true for any population, regardless whether it is normal or not.

Estimate of \sigma_{\bar{Y}} = \sigma/\sqrt{n} = s/\sqrt{n}.

For a normal distribution only, S.E.(s) = \sigma_s = \sigma/\sqrt{2n} and the estimate of \sigma_s = \hat{\sigma}_s = s/\sqrt{2n}.

Central Limit Theorem: For a large n, regardless of the underlying distribution of the individual observations is \bar{Y} \sim N(\mu, \sigma^2/n).

ESTIMATION OF μ

Assumptions:

1. Observations are normally distributed: $Y \sim N(\mu, \sigma^2)$.
2. σ^2 is known.

The point estimate for Y_1, Y_2, \dots, Y_n is $\hat{\mu} = \bar{Y}$.

A more informative way of writing this point estimate is $\bar{Y} \pm s_{\bar{Y}}$, where $s_{\bar{Y}} = s/\sqrt{n}$.

This expression tells us something about the precision of the estimate.

Thus, it is not sufficient to give only \bar{Y} , but also \pm S.E. or standard deviation of \bar{Y} .

Interval Estimation:

$$L_1 \leq \mu \leq L_2$$

L_1 is the lower confidence limit and L_2 is the upper confidence limit.

L_1 and L_2 are functions of

- a) \bar{Y} (the point estimate)
- b) σ (known)
- c) n (sample size)
- d) α (confidence placed in the statement)

How to Obtain Confidence Limits L_1 and L_2 :

Situation 1: $Y \sim N(\mu, \sigma^2)$, where μ is unknown, but σ^2 is known.

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Convert this normal distribution to the standard normal distribution:

$$\bar{X} = (\bar{Y} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$$

$$\Pr[-z \leq \bar{X} \leq z] = 1 - \alpha.$$

$1 - \alpha$ is the coefficient of confidence.

Choose a confidence coefficient, $1 - \alpha$, e.g., 0.95 (95% confidence interval for μ):

$$z_{1-\alpha/2} = z_{.975} = 1.96 \text{ and}$$

$$z_{\alpha/2} = z_{.025} = -1.96.$$

$$\Pr[-1.96 \leq (\bar{Y} - \mu)/(\sigma/\sqrt{n}) \leq 1.96] =$$

$$\Pr[-1.96(\sigma/\sqrt{n}) \leq \bar{Y} - \mu \leq 1.96(\sigma/\sqrt{n})] =$$

$$\Pr[\bar{Y} - 1.96(\sigma/\sqrt{n}) \leq \mu \leq \bar{Y} + 1.96(\sigma/\sqrt{n})] = 0.95$$

In general, $(\bar{Y} - z_{1-\alpha/2}(\sigma/\sqrt{n}), \bar{Y} + z_{1-\alpha/2}(\sigma/\sqrt{n}))$ is called the $(1 - \alpha)100\%$ confidence interval for μ .

In this context, both μ and the length, ℓ , are fixed.

\bar{Y} changes with each new sample – hence, only the location of the C.I. will change.

The Meaning or Interpretation of a Confidence Interval:

With repeated experiments, for each sample a different L_1 and L_2 will be computed because \bar{Y} will be different with each sample; $(1 - \alpha)100\%$ of the confidence intervals will include μ and $\alpha 100\%$ will not.

Thus, we are $(1 - \alpha)100\%$ confident that the interval contains μ .

How to decrease the length of the confidence interval for a fixed α :

$$\text{distance} = \ell = 2(z_{1-\alpha/2}(\sigma/\sqrt{n})).$$

1. Increase the n , sample size, or choose a very large n ($n = \infty$) so that the quantity approaches zero and \bar{Y} approaches the true μ .
2. For a fixed n , reduce $z_{1-\alpha/2}$, reduce $1 - \alpha$ and increase α so that we will be less confident in the statement.

This formula illustrates a conflicting problem in that we want high level of confidence but an interval that is not too large.

Traditional levels of confidence are .90, .95, and .99.

It is crucially important to let others know what level of confidence we have chosen.

Sample Size Determination When σ Is Known:

Choosing n based on the length of C.I. where $2z_{1-\alpha/2}(\sigma/\sqrt{n}) = \ell$,

$$n = [(2z_{1-\alpha/2})\sigma / \ell]^2$$

Take the next larger integer so that C.I. will not exceed ℓ .

Situation 2: Confidence Interval for μ when σ^2 is unknown.

In the formula for the standardized normal variate (z-score) above, replace σ^2 by s^2 ,
i.e., σ/\sqrt{n} by s/\sqrt{n} .

Then, $(\bar{Y} - \mu)/(s/\sqrt{n}) \sim t_{[n-1]}$ where $t_{[n-1]}$ is Student's t-distribution with $n - 1$ degrees of freedom (d.f.).

When σ is replaced by s , we can get a standardized normal deviate, $(\bar{Y} - \mu)/(s/\sqrt{n})$, but this no longer follows a normal distribution. The $t_{[n-1]}$ distribution is bell-shaped and symmetrical but there are some differences. For each d.f., there is a different distribution, where d.f. refers to the d.f. that s is based on.

$$\Pr[-t_{1-\alpha/2[n-1]} \leq (\bar{Y} - \mu)/(s/\sqrt{n}) \leq t_{1-\alpha/2[n-1]}] = 1 - \alpha.$$

Implies

$$L_1 = \bar{Y} - t_{1-\alpha/2[n-1]} * s/\sqrt{n}$$

$$L_2 = \bar{Y} + t_{1-\alpha/2[n-1]} * s/\sqrt{n}$$

gives $(1 - \alpha)100\%$ confidence interval.

μ is fixed; L_1 and L_2 are variable so that $(1 - \alpha)100\%$ of the intervals will contain μ .

Both \bar{Y} and s change with each new sample – both location and length of the C.I. will change.

In addition, $t_{1-\alpha/2}$ will always be larger than the corresponding $z_{1-\alpha/2}$ – on the average C.I.'s will be larger when σ is unknown than when σ is known.

Example:

$$\bar{Y} = 4.537$$

$$s = .3812$$

$$n = 15$$

$$\alpha = 0.05; 1 - \alpha = 0.95$$

$t_{.975[14]} = 2.145$ from table 5.

$$\Pr[\bar{Y} - t_{1-\alpha/2[n-1]} * s/\sqrt{n} \leq \mu \leq \bar{Y} + t_{1-\alpha/2[n-1]} * s/\sqrt{n}] = 1 - \alpha.$$

$$4.537 - 2.145 * (.3812/\sqrt{15}) \leq \mu \leq 4.537 + 2.145 * (.3812/\sqrt{15})$$

$$4.326 \leq \mu \leq 4.748$$

Conclusion: We are 95% confident that the true μ (which is unknown) lies between 4.326 and 4.748.

Length of C.I. = $L_2 - L_1 = \ell = 2t_{1-\alpha/2[n-1]} * s/\sqrt{n}$ is a random variable which depends on the sample through s .

Sample Size Determination When σ Is Unknown:

We have to have some information about s to determine appropriate sample size, n .

Example:

Guess: $s = 2$

Assume: This quantity is also the value for σ from the sample, which is yet to be taken!

Choose: $1 - \alpha = 0.95$

Want: $\ell = 3$

From σ^2 -known case:

$$n^* = [(2 * 1.96\sigma)/\ell]^2 = [(2 * 1.96*2)/3]^2 = 6.83 \approx 7$$

Equate $(2z_{1-\alpha/2}\sigma)/\sqrt{n^*} = (2t_{1-\alpha/2[n^*-1]}\sigma)/\sqrt{n}$:

$$z_{1-\alpha/2}/\sqrt{n^*} = t_{1-\alpha/2[n^*-1]}/\sqrt{n}$$

Solve for n :

$$n = n^*[t_{1-\alpha/2[n^*-1]}/z_{1-\alpha/2}]^2$$

Remember many assumptions have been made. This does not guarantee that ℓ will be equal to the value we want in the actual sample, but this gives us some idea of what n should be. This step is often overlooked.

ESTIMATION OF σ^2 AND σ

Point Estimation: $\hat{\sigma}^2 = s^2$ and $\hat{\sigma} = s$.

$$\text{S.E.}(s^2) = \sqrt{2} * \sigma^2 / \sqrt{(n - 1)}$$

$$\text{S.E.}(s) = \sigma / \sqrt{2n}$$

$$\text{Estimated S.E.}(s^2) = \sqrt{2} * s^2 / \sqrt{(n-1)}$$

$$\text{Estimated S.E.}(s) = s / \sqrt{2n}$$

The estimated S.E. is a measure of the variability or dispersion from sample to sample. This S.E. is usually large unless we take many observations.

Interval Estimation:

$$L_1 \leq s^2 \leq L_2$$

L_1 and L_2 are functions of

1. s^2
2. n
3. confidence or $1 - \alpha$

Does not depend on \bar{Y} – is independent of the mean.

$$(n-1)s^2/\sigma^2 \sim \chi^2_{[n-1]}$$

We have to know the distribution of s^2 to determine the confidence limits.

There is a different χ^2 distribution for each d.f. or n .

χ^2 is not a symmetric distribution, but as d.f. increase the shape becomes more symmetric.

Example to find percentage points in table 6:

$$\alpha = .05$$

$$\alpha/2 = .025 \quad \chi^2_{.025[9]} = 2.700$$

$$1 - \alpha/2 = .975 \quad \chi^2_{.975[9]} = 19.023$$

$$\Pr[\chi^2_{\alpha/2[n-1]} \leq (n-1)s^2/\sigma^2 \leq \chi^2_{1-\alpha/2[n-1]}] = 1 - \alpha$$

$$\Pr[(n-1)s^2/\chi^2_{1-\alpha/2[n-1]} \leq \sigma^2 \leq (n-1)s^2/\chi^2_{\alpha/2[n-1]}] = 1 - \alpha$$

$$L_1 = (n-1)s^2/\chi^2_{1-\alpha/2[n-1]}$$

$$L_2 = (n-1)s^2/\chi^2_{\alpha/2[n-1]}$$

If we do 100 such limits, 95 will include σ^2 and 5 will not, given $\alpha = .05$.

L_1 and L_2 are always positive.

As confidence increases, the interval will get larger.

To reduce interval, lower confidence coefficient or increase n .

Confidence Interval for σ :

$$L'_1 \leq s \leq L'_2$$

$$L'_1 = \sqrt{L_1}$$

$$L'_2 = \sqrt{L_2}$$

HYPOTHESIS TESTING ONE-SAMPLE INFERENCE

Example:

Research Question: Do babies of mothers with low SES have lower birth weights than average?

Birth weight of infants
National ave.: $\mu = 120$ oz. = μ_0
Low SES ave: $\bar{Y} = 115$ oz.
 $\sigma = 25$ oz.
 $n = 100$

Assume: All parameters are known, except $\mu_{\mathcal{F}}$.

$$\bar{Y} \sim N(\mu_{\mathcal{F}}, 6.25).$$

Distribution of sample means is important.

$$\begin{aligned}\sigma_{\bar{Y}}^2 &= 25^2/n = 625/100 = 6.25 \\ \sigma_{\bar{Y}} &= 2.5 \\ z_s &= (\bar{Y} - \mu)/2.5 \sim N(0, 1)\end{aligned}$$

z_s is a sample statistic, which is associated with a probability.

We will determine whether μ_{SES} or $\mu_{\mathcal{F}}$ is either equal to the national average or it is not.

$$\begin{aligned}\text{Test } H_0: & \mu_{\mathcal{F}} = \mu_0 \text{ with } \sigma = \sigma_0 \\ & \text{versus} \\ H_1: & \mu_{\mathcal{F}} \neq \mu_0 \text{ with } \sigma = \sigma_0\end{aligned}$$

The hypothesis is always formulated about parameters.

H_0 designates the **null hypothesis** and H_1 the **alternative hypothesis**.

Based on sample statistics we will choose which is the true situation.

Our reasons for choice will be based on how unlikely our sample result could have been obtained from a particular population, given the null hypothesis is true.

In this example we will use the sample statistic z_s as the basis for our decision.

In making this choice (decision), we may make errors.

We always want to minimize the chances of erroneous decisions even though we can never know the truth.

		TRUE STATE OF NATURE	
		H_0 true	H_1 true
DECISION IN FAVOR OF	H_0	Correct	Type II Error
	H_1	Type I Error	Correct

$\Pr[\text{Reject } H_0 \mid H_0 \text{ is correct}] = \Pr[\text{Type I Error}] = \alpha = \text{the probability of making a Type I error.}$

$\alpha * 100\% = \text{significance level.}$

We want α to be small to emphasize rejecting H_0 when we should do so, given the nature of experiments and observational studies.

$\Pr[\text{Not rejecting } H_0 \mid H_1 \text{ is correct}] = \Pr[\text{“Accepting” } H_0 \mid H_1 \text{ is correct}] = \Pr[\text{Type II Error}] = \beta$

Before we collect data, we should choose α , which will influence the size of β .

$1 - \beta = \text{power of test.}$

This is the power of discriminating between H_0 and H_1 .

Power depends on

1. α ,
2. H_1 and H_0 , i.e., $\Delta = \mu_{SES} - \mu_0$ and
3. n .

A typical value for power is .8.

Hypothesis Testing Procedure

1. Set up H_0 . (Want to reject this)
2. Set up H_1 .
3. Choose α .
4. Determine n .
5. Obtain data.
6. Compute test statistic in terms of parameters under H_0 .
7. Perform test: Does the test statistic fall into the **critical region**?
Yes: Reject H_0 . **No:** Fail to reject H_0 .

One-Tailed Hypothesis Tests About μ When σ Is Known

We define a **one-tailed test** as a test in which the parameter of interest (e.g., μ) under H_1 can only be greater than or less than the value under H_0 .

Example:

$$H_0: \mu = 120 \text{ with } \sigma_{\bar{Y}}^2 = 6.25.$$

$$H_1: \mu < 120 \text{ with } \sigma_{\bar{Y}}^2 = 6.25.$$

$$\alpha = .05.$$

$$n = 100.$$

$$\text{From data } \bar{Y} = 115.$$

$$z_s = (115 - 120)/2.5 = -2.0.$$

Critical region: $(-\infty, -1.645)$ to the left of which point lies 5% of the population.

$$-2.0 < -1.645 \Rightarrow \text{Reject } H_0.$$

In general, if we wish to test the hypothesis $H_0: \mu = \mu_0 \mid \sigma = \sigma_0$ vs. $H_1: \mu < \mu_0 \mid \sigma = \sigma_0$ with a significance level of α , then the most powerful test is based on \bar{Y} :

If $\bar{Y} < \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}$, then reject H_0 .

If $\bar{Y} \geq \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}$, do not reject H_0 .

In the statistical sense, *powerful* means having the ability to discriminate between the null and alternative hypotheses.

Similarly, if $H_1: \mu > \mu_0$, with $\sigma = \sigma_0$, with a significance level of α , then based on our sample statistic \bar{Y} :

If $\bar{Y} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$, then reject H_0 .

If $\bar{Y} \leq \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$, do not reject H_0 .

The P-Value for a One-Sided Z-Test

The **p-value** is defined as the level of significance at which we would be unable to choose between the two hypotheses, given our data.

On the **standard normal table**, we look up the probability for obtaining a value more extreme than the test statistic z_s , which we have calculated based on our data; this probability is the p-value.

Thus, the **p-value** is the probability of obtaining this result by chance alone if H_0 were true.

The p-value tells exactly how statistically significant our results are.

Statistical significance is the probability associated with the sample statistic computed from the data.

Statistical significance has absolutely nothing to do with the scientific or clinical importance of results.

Two-Tailed Hypothesis Tests About μ When σ Is Known

We do not always have prior knowledge of direction in which the alternative mean will lie.

A **two-tailed test** is a test in which the values of the parameter under H_1 can be either less than or greater than the value of the parameter under H_0 .

We wish to test

$H_0: \mu = \mu_0$ with $\sigma = \sigma_0$ versus

$H_1: \mu \neq \mu_0$ with $\sigma = \sigma_0$ at $\alpha = .05$.

If $\bar{Y} < \mu_0 - z_{1-\alpha/2} * \sigma / \sqrt{n}$ or $\bar{Y} > \mu_0 + z_{1-\alpha/2} * \sigma / \sqrt{n}$, then we reject H_0 .

If \bar{Y} lies between these values, we do not reject H_0 .

The P-Value for a Two-Sided Z-Test

The p-value for a two-sided test is calculated by multiplying the standard normal probability associated with our test statistic z_s by 2.

It is always more difficult to reject a two-sided test than a similar one-sided test.

What about β ?

A lower α (Type I Error) affects the acceptance and rejection of H_0 ; therefore, with a smaller α we will have a larger β .

The **critical region** or the **region of rejection** is determined by

1. choice of α
2. form of H_1 (e.g., one- or two-sided)
3. $\Delta = |\mu_1 - \mu_0|$ and
4. distribution of the test statistic, e.g., normal, t, χ^2 .

We must always qualify the rejection of a hypothesis by giving the level of significance.

Probability of Type II Error depends on

1. α : If α increases, β decreases; if decrease α , increase β .
2. H_1 in relation to H_0 : If $\Delta = |\mu_0 - \mu_1|$ decreases, β increases because, as two means become closer, it is more difficult to distinguish between the two hypotheses.
3. σ : If σ increases, β increases. With small σ , distribution is more peaked and β smaller.
4. n : If n increases, β decreases because the standard error gets smaller.

Power of the test, $1-\beta$, is the power of discriminating between H_0 and H_1 .

We want $1-\beta$ to be “large” because $1-\beta$ is the probability of correctly rejecting H_0 .

Automatically, as β decreases, $1-\beta$ increases.

Power is very dependent on sample size, n .

Example:

$H_0: \mu_0 = 120$ vs. $H_1: \mu_1 < 115$ with $\sigma = 25$ and $\alpha = .05$.

Rejection Region:

$$z_s = (\bar{Y} - 120)/(25/\sqrt{n}) \leq -1.645 \text{ or} \\ \bar{Y} \leq 120 - 1.645*25/\sqrt{n}$$

$$1-\beta = \Pr[\bar{Y} \leq 120 - 1.645*25/\sqrt{n} \mid H_1]$$

$$= \Pr[(\bar{Y} - 115) / (25/\sqrt{n}) \leq (120 - (1.645*25/\sqrt{n}) - 115) / 25/\sqrt{n} \mid H_1]$$

$$= \Pr[z \leq (5 - 1.645*25/\sqrt{n})/25/\sqrt{n} \mid H_1] \Rightarrow 1-\beta = \Pr[z < z_\alpha + (\Delta\sqrt{n})/\sigma \mid H_1]$$

A good level for $1-\beta$ is 0.8.

Depending on circumstances, power may be increased to 0.99.

Often in unplanned studies, power is as low as 0.2 or less.

$$\text{Power} = 1-\beta = \Pr[z \leq z_\alpha + (\Delta\sqrt{n})/\sigma] = \Phi(z_{1-\beta}) \\ = \text{PROBNORM}(z_{1-\beta})$$

Testing for the Mean of a Normal Population with σ^2 Unknown

Types of Hypotheses:

1. Two-sided test – $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$
2. One-sided test – $H_0: \mu \leq \mu_0$ vs. $H_1: \mu > \mu_0$
3. One-sided test – $H_0: \mu \geq \mu_0$ vs. $H_1: \mu < \mu_0$
4. Simple hypothesis – $H_0: \mu = \mu_0$ vs. $H_1: \mu = \mu_1$

Test statistic:

$$t_s = (\bar{Y} - \mu_0) / s / \sqrt{n}$$

follows a t-distribution with $n-1$ d.f.

Region of Rejection

If values are in this range, reject H_0 .

1. $|t_s| > t_{1-\alpha/2[n-1]}$ for a two-sided test.
2. $t_s > t_{1-\alpha[n-1]}$ for a one-sided upper-tail test.
3. $t_s < t_{\alpha[n-1]}$ for a one-sided lower-tail test.
4. For simple hypothesis tests
 $\mu_1 > \mu_0: t_s > t_{1-\alpha[n-1]}$ or $\mu_1 < \mu_0: t_s < t_{\alpha[n-1]}$.

Thus, $t_s = (\bar{Y} - \mu_0) / s / \sqrt{n}$ is the test statistic for testing $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$ or any of the other examples listed above.

The P-Value for a One- or Two-Sided T-Test

On **Student's t table for n-1 d.f.**, we look up the probability for obtaining a value more extreme than the test statistic t_s , which we have calculated based on our data; this probability is the p-value. Thus, the **p-value** is the probability of obtaining this result by chance alone if H_0 were true.

The p-value for a one-sided test is the probability associated with our test statistic t_s .

The p-value for a two-sided test is calculated by multiplying the probability associated with our test statistic t_s by 2.

Determination of Sample Size and Power When σ Is Unknown

Sample size has an important impact on the critical value.

For the same sample size, it is harder to reject H_0 , when the σ is unknown: $|t_s| > |z_s|$.

We can obtain **Power Curves**, also called **OC (Operating Characteristic) Curves**, which relate power to effect size for a given sample size, to show that power changes with each situation, as determined by two-sided or one-sided tests and α .

For a sample size of n , power will change depending on whether the test is one- or two-sided.

For a sample size of n , power will change for each alternative mean (μ_1) chosen.

Sample Size Determination

1. Determine the type of test – one-sided or two-sided.
2. Choose α .
3. Choose $1-\beta$ or power to reject H_0 when we should.
4. Determine what **difference in standard deviation units** (i.e., **effect size**) you want to detect, provided such a difference exists: $\Delta = |\mu_0 - \mu_1| / \sigma$

5. For one-sided test,
$$n = [\sigma(z_{1-\beta} + z_{1-\alpha}) / |\mu_0 - \mu_1|]^2$$

For two-sided test,
$$n = [\sigma(z_{1-\beta} + z_{1-\alpha/2}) / |\mu_0 - \mu_1|]^2$$

Sample size is exceedingly sensitive to the difference between the means.

The difference between means should be that difference, which is scientifically, biologically, or clinically important, in the judgement of the researcher.

When no information is available, a pilot study may be conducted to get some idea of differences that can be obtained in a particular experimental or clinical situation.

Data from many studies fail to show statistically significant and clinically important differences between treatment groups, when in fact these exist – solely due to inadequate sample size.

Trivial differences have also been found statistically significant and misleadingly reported as scientifically important – solely because of extremely large sample sizes.

Unfortunately, decisions based on either of the foregoing scenarios can involve enormous amounts of money or extremely large intangible costs.

Relationship Between C.I.'s and Hypothesis Testing

CI's give a range of values within which μ is likely to fall.

CI's do not use a p-value; we get this information from hypothesis testing.

Sample size is as important for CI's as for hypothesis testing.

General result: If H_0 is rejected, then the corresponding confidence interval does not contain the parameter under H_0 . The one-to-one relationship between a confidence interval and the corresponding hypothesis test is easiest to represent with the two-sided case.

For completeness it is a good practice to report enough information that both CI's and p-values are obvious to anyone reading the report.

Testing Hypotheses About σ^2

Most frequently used hypothesis test is a two-sided one.

Test Statistic: $X_s^2 = (n-1)s^2/\sigma^2$.

Test at $\alpha*100\%$ level.

1. **Two-sided test** to choose between:

$$H_0: \sigma^2 = \sigma_0^2 \text{ vs. } H_1: \sigma^2 \neq \sigma_0^2$$

Critical Regions:

$$(0, \chi^2_{\alpha/2[n-1]}) \text{ \underline{and} } (\chi^2_{1-\alpha/2[n-1]}, +\infty)$$

$$\text{If } X_s^2 < \chi^2_{\alpha/2[n-1]} \text{ \underline{or} } X_s^2 > \chi^2_{1-\alpha/2[n-1]} \Rightarrow \text{Reject } H_0.$$

2. **One-sided test** to choose between:

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 = \sigma_1^2$$

- a) $\sigma_1^2 > \sigma_0^2$

Critical Region: $(\chi^2_{1-\alpha[n-1]}, +\infty)$

$$\text{If } X_s^2 > \chi^2_{1-\alpha[n-1]} \Rightarrow \text{Reject } H_0.$$

- b) $\sigma_1^2 < \sigma_0^2$

Critical Region: $(0, \chi^2_{\alpha[n-1]})$

$$\text{If } X_s^2 < \chi^2_{\alpha[n-1]} \Rightarrow \text{Reject } H_0.$$

3. **One-sided test** to choose between:

$$H_0: \sigma^2 \leq \sigma_0^2 \text{ vs. } H_1: \sigma^2 > \sigma_0^2$$

One-sided upper-tail test, same as *2a*.

4. **One-sided test** to choose between:

$$H_0: \sigma \geq \sigma_0^2 \text{ vs. } H_1: \sigma^2 < \sigma_0^2$$

One-sided lower-tail test same as *2b*.

The P-Value for a One- or Two-Sided Chi-Square Test

For a one-sided test, p-value is the probability associated with X_s^2 for $n-1$ d.f.

For the two-sided test multiply the associated probability for X_s^2 with $n-1$ d.f. by 2.

Sample Size Determination for Hypothesis Tests About σ^2

1. Determine the type of test:
two-sided or one-sided upper-tail, one-sided lower-tail.
2. Choose α .
3. Choose $1-\beta$.
4. Decide what ratio $\lambda = \sigma_1/\sigma_0$ of the standard deviations (i.e., effect size) you want to detect.