AP Statistics

03 May 2023

Revision: 435

Aziz Manva

azizmanva@gmail.com

Table of contents

1. Statistics

1.1 Descriptive Statistics

1.1: Mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

- You add all the data points, and then divide by the number of data points.

1.2: Deviation from the Mean

$$x_i - \bar{x}$$

1.3: Standard Deviation

The standard deviation is the square root of the average squared deviation from the mean.

$$Sample\ Standard\ Deviation = s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$Population\ Standard\ Deviation = \sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Example 1.4

Find the sample standard deviation for:

$$4,8,9$$

$$\bar{x} = \frac{4+8+9}{3} = \frac{21}{3} = 7$$
$$(x_1 - \bar{x})^2 = (4-7)^2 = (-3)^2 = 9$$
$$(x_2 - \bar{x})^2 = (8-7)^2 = 1^2 = 1$$
$$(x_3 - \bar{x})^2 = (9-7)^2 = 2^2 = 4$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{9+1+4}{3-1}} = \sqrt{\frac{14}{2}} = \sqrt{7}$$

1.2 Random Variables

A. Basics

1.5: Random Variable

If you conduct an experiment and assign the value to a variable, that variable is a random variable.

- Random variable are usually assigned capital letters X, Y, Z
- The values that random variables can take are usually assigned small letters x, y, z.

1.6: Expected Value

$$Mean = E[X] = \mu = \sum_x x \cdot p(x)$$

- μ is the Greek letter mu.
- $E[X]$ is an operator (or function). It is not $E \cdot X$

1.7: Variance

$$Variance = \sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 \cdot p(x)$$

1.8: Standard Deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_x (x-\mu)^2 P(x)}$$

Example 1.9

A. What is the probability that the number of heads is less than or equal to 1?

x	0	1	2
$P(X=x)=p(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$P(X \le 1) = P(X=0) + P(X=1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

Example 1.10

I toss a fair coin twice, and count the number of heads X.

A. Calculate the probability distribution function for X.
B. Calculate the mean of X.
C. Calculate the variance of X.
D. Calculate the standard deviation of X.

x	0	1	2	
$P(X=x)=p(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	
$xp(x)$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\mu = 1$
$x-\mu$	-1	0	1	
$(x-\mu)^2$	1	0	1	
$(x-\mu)^2 p(x)$	$\frac{1}{4}$	0	$\frac{1}{4}$	$\sigma^2 = \frac{1}{2}$

B. Properties of Probability

1.11: Range of Probability

$$0 \le P(X=x) \le 1, for\ all\ x$$

1.12: Sum of Probabilities

Sum of all the probabilities of a discrete probability distribution is one.

Example 1.13

I toss a fair coin twice, and count the number of heads X. Verify that the sum of probabilities is 1.

x	0	1	2	Sum
$P(X=x)=p(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

1.14: Cumulative Probabilities

$$F(X_0)=P(X\le x_0)=\sum_{X\le X_0}P(X)$$

- Cumulative probabilities are non decreasing.

Example 1.15

I toss a fair coin twice, and count the number of heads X. Write the cumulative probability distribution.

x	0	1	2
$P(X=x)=p(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$F(x)$	$\frac{1}{4}$	$\frac{3}{4}$	1

C. Symmetric Distribution

1.16: Symmetric Distribution

A symmetric distribution will a vertical line of symmetry.

Example 1.17

I toss a fair coin thrice, and count the number of heads X. Is the probability distribution symmetric?

x	0	1	2	3
$P(X=x)=p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$Yes$$

1.18: Symmetric Distribution

In a symmetric distribution:

$$Mean=Median=Mode$$

D. Transformations of Random Variables

1.19: Transformations

$$E[cX]=cE[X]$$
$$Var[cX]=c^2Var[X]$$

$$SD[cX] = |c|SD[X]$$

Example 1.20

A. X is a random variable with mean 5 and standard deviation 3. Determine the mean and standard deviation of 3X.

B. The number of monthly international phone calls that Anaya makes is a random variable with mean 12 and variance 2. The cost of each phone call is $5. Find the mean and the standard deviation of the monthly cost of the international phone calls that Anaya makes.

Part A

$$Mean[3X] = E[3X] = 3E[X] = 3(5) = 15$$
$$SD[3X] = 3SD[X] = 3(3) = 9$$

Part B

Let the random variable

$$X = No. of\ monthly\ international\ phone\ calls$$

We want to find:

$$E[5X] = 5E[X] = 5 \cdot 12 = 60$$
$$SD[5X] = 5SD[X] = 5\sqrt{2}$$

1.21: Constant Values

$$E[c] = c$$
$$Var[c] = 0$$

1.22: Linearity of Expectation

$$E[X + Y] = E[X] + E[Y]$$
$$Var[X + Y] = Var[X] + Var[Y]\ if\ X\ and\ Y\ are\ independent$$

1.23: Adding a Constant

$$E[a + X] = a + E[X]$$
$$Var[a + X] = Var[X]$$

$$E[a + X] = E[a] + E[X] = a + E[X]$$
$$Var[a + X] = Var[a] + Var[X] = Var[X]$$

Example 1.24

1.25: Linear Combinations of Random Variables

$$E[aX + bY] = aE[X] + bE[Y]$$

$$E[aX + bY] = E[aX] + E[bY] = aE[X] + bE[Y]$$

1.26: Linear Combinations of Random Variables

If X and Y are independent:

$$Var[aX + bY] = a^2Var[X] + b^2Var[Y]$$

$$Var[aX + bY] = Var[aX] + Var[bY] = a^2Var[X] + b^2Var[Y]$$

Example 1.27

Show that $Var[aX - bY] = a^2Var[X] + b^2Var[Y]$

Use a change of variable.

$$Var[aX + (-b)Y] = a^2Var[X] + (-b)^2Var[Y] = a^2Var[X] + b^2Var[Y]$$

Example 1.28

The runs scored by a cricketer in an innings of a test match are a random variable X with mean 50 and standard deviation 10. The runs in two different innings are independent of each other.

A. Determine an expression for the random variable Y that tracks the average score in a test match with two innings.
B. Show that $E[Y] = E[X]$
C. Show that $SD[Y] < SD[X]$

Part A

Let X_1 and X_2 be random variables that follow the probability distribution X. We write

$$X_1, X_2 \sim X(\mu = 50, \sigma = 10)$$

Then the random variable Y is defined as:

$$Y = \frac{X_1 + X_2}{2}$$

Part B

$$E[Y] = E\left[\frac{X_1 + X_2}{2}\right] = E\left[\frac{1}{2}X_1 + \frac{1}{2}X_2\right] = \frac{1}{2}E[X_1] + \frac{1}{2}E[X_2] = \frac{1}{2}E[X] + \frac{1}{2}E[X] = E[X]$$

Part C

$$Var[Y] = Var\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{4}Var[X_1] + \frac{1}{4}Var[X_2] = \frac{1}{4}Var[X] + \frac{1}{4}Var[X] = \frac{1}{2}Var[X]$$

$$Var[Y] = \frac{1}{2}Var[X]$$

$$SD[Y] = \sqrt{Var[Y]} = \sqrt{\frac{1}{2}Var[X]} = \frac{1}{\sqrt{2}}SD[Y]$$

1.3 Distributions

A. Bernoulli Distributions

Now we turn our attention to specific distribution that are useful in statistics. Since these distributions are used very frequently, they have names, and their properties are very important, and should be memorized.

1.29: Bernoulli Distribution

A random variable X follows a Bernoulli distribution if

- it has exactly two outcomes: success, and failure.
- Probability of success $= p$

- Traditionally, success is traditionally assigned a value of 1, and failure is assigned a value of 0.
- Success does not have to mean success in the common-sense understanding. For example, the death of a patient in a hospital can be termed a "success".

Example 1.30

I have an urn with 3 blue balls and some red balls. The total number of balls is 8. I draw a ball from the urn, and assign the random variable X as 1 if the ball is blue, and 0 otherwise.

A. Determine the distribution that X follows.
B. Write it in a probability distribution.
C. Calculate the mean and the variance of X.

$$X \sim Bernouli$$
$$Success = P(Blue) = \frac{3}{8}$$
$$Failure = P(Red) = \frac{5}{8}$$

	Failure $x = 0$	$Success$ $x = 1$	
$P(X = x)$	$\frac{5}{8}$	$\frac{3}{8}$	
$xp(x)$	0	$\frac{3}{8}$	$\mu = \frac{3}{8}$
$(x - \mu)$	$-\frac{3}{8}$	$\frac{5}{8}$	
$(x - \mu)^2$	$\frac{9}{64}$	$\frac{25}{64}$	
$(x - \mu)^2 p(x)$	$\frac{45}{512}$	$\frac{75}{512}$	$\sigma^2 = \frac{120}{512} = \frac{15}{64}$

Example 1.31

X follows a Bernoulli distribution with probability of success p.

A. Determine the probability of failure, and write the probability distribution of X as a table.
B. Determine the mean, variance and standard deviation of X.

Part A

Let probability of failure be q.

$$p + q = 1 \Rightarrow q = 1 - p$$

x	0	1
$P(X = x)$	$1 - p$	p
$x - \mu$	$-p$	$1 - p$

Part B

Mean

$$\sum_x xP(x) = (0)(1 - p) + (1)(p) = p$$

Variance

$$Variance = \sigma^2 = \sum_x (x - \mu)^2 P(x)$$
$$= (-p)^2(1 - p) + (1 - p)^2(p)$$
$$= p^2 - p^3 + (1 - 2p + p^2)(p)$$
$$= p^2 - p^3 + p - 2p^2 + p^3$$
$$= p^2 + p - 2p^2$$
$$= -p^2 + p$$
$$= p(1 - p)$$

Standard Deviation

$$Standard\ Deviation = \sigma = \sqrt{\sigma^2} = \sqrt{p(1 - p)}$$

1.32: Bernoulli Distribution: Summary

If X follows a Bernoulli distribution with parameter p, we write

$$X \sim Bernoulli(p)$$

	$Mean = \mu$	$Variance = \sigma^2$	$SD = \sigma$
X	p	$p(1-p)$	$\sqrt{p(1-p)}$

Example 1.33

I have an urn with 3 blue balls and some red balls. The total number of balls is 8. I draw a ball from the urn, and assign the random variable X as 1 if the ball is blue, and 0 otherwise.

A. Determine the distribution that X follows.
B. Write it in a probability distribution.
C. Calculate the mean and the variance of X.

$$p = \frac{3}{8} \Rightarrow 1 - \frac{5}{8}$$
$$\mu = p = \frac{3}{8}$$
$$\sigma^2 = p(1-p) = \frac{3}{8}\left(\frac{5}{8}\right) = \frac{15}{64}$$

Example 1.34

A student attempts to solve a statistics question. His probability of success is known to be 0.4. If we consider a success as 1, and a failure as zero, what is the standard deviation of the outcome?

$$p = 0.4 \Rightarrow SD = \sqrt{0.4 \cdot 0.6} = \sqrt{0.24} = \sqrt{\frac{24}{100}} = \frac{2\sqrt{6}}{10}$$

1.35: Constant Property: Mean and Variance

$$E[aX] = aE[X]$$
$$Var[aX] = a^2 Var[X]$$
$$SD[aX] = |a| SD[X]$$

Example 1.36

A student attempts to solve a statistics question. His probability of success is known to be 0.4. The student will get awarded 10 marks on success, and 0 marks on failure. Calculate the expected value and the standard deviation of his marks.

Let X be a random variable that has the outcome for the statistics question.

$$p = 0.4, Success \Rightarrow X = 1, Failure \Rightarrow X = 0$$

$$E[10X] = 10E[X] = 10 \times 0.4 = 4$$
$$SD[10X] = 10SD[X] = 10 \cdot \frac{2\sqrt{6}}{10} = 2\sqrt{6}$$

B. Binomial Distribution

1.37: Binomial Distribution

A random variable X follows a binomial distribution if it consists of n independent, identically distributed Bernoulli distributions.

- The random variable X takes the value of the number of successes in the n trials.

Example 1.38

Determine the mean and the variance of a Binomial distribution with n trials and probability of success p.

Let

$$X \sim Binomial(n, p)$$

Mean

Note that X consists of n independent, identically distributed Bernoulli distributions:

$$E[X] = E[X_1 + X_2 + \cdots + X_n]$$

Using the property of expectation:

$$= E[X_1] + E[X_2] + \cdots + E[X_n]$$

But note that each Bernoulli distribution has $success\ parameter = mean = p$:

$$= \underbrace{p + p + \cdots + p}_{n\ times} = np$$

Variance

Note that X consists of n independent, identically distributed Bernoulli distributions:

$$Var[X] = Var[X_1 + X_2 + \cdots + X_n]$$

Using the property of variance:

$$= Var[X_1] + Var[X_2] + \cdots + Var[X_n]$$

But note that each Bernoulli distribution has $variance = p(1-p)$:

$$= \underbrace{p(1-p) + p(1-p) + \cdots + p(1-p)}_{n\ times} = np(1-p)$$

Standard Deviation

$$Standard\ Deviation = \sigma = \sqrt{\sigma^2} = \sqrt{np(1-p)}$$

1.39: Binomial Distribution: Summary

If X follows a Binomial distribution with n trials, and probability of success p,

$$X \sim Binomial(n, p)$$

	$Mean = \mu$	$Variance = \sigma^2$	$SD = \sigma$
X	np	$np(1-p)$	$\sqrt{np(1-p)}$

Example 1.40

Ralph tosses a weighted coin that lands on heads with probability 0.3. Calculate the expected value and the variance of the number of heads in 5 tosses.

Let the number of heads in 5 tosses be X.

$$X \sim Binomial(5, 0.3)$$

$$E[X] = np = 5(0.3) = 1.5$$
$$Var[X] = np(1-p) = 1.5 \times 0.7 = 1.05$$

C. Calculation of Probabilities

1.41: Probability Distribution Function

The probability of x successes for a binomial distribution with n trials and probability of success p is:

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

Example 1.42

With a fair coin, what is the probability of getting two heads in seven tosses? three heads in five tosses? two heads in eight tosses? three heads in ten tosses?

$$P(X=2)=\binom{n}{x}p^x(1-p)^{n-x}$$

Example 1.43

I visit the gym everyday. At the gym, I am equally likely to run on the treadmill or do weight training when I visit the gym (which I do every day). What is the probability that I run on the treadmill exactly four days in the second week of February?

D. Geometric Distribution

1.44: Geometric Distribution

If you have independent, identical Bernoulli trials being performed with probability of success p, then the random variable Y that counts the number of trials till the first success has a geometric distribution

1.45: PDF of Geometric Distribution

$$P(Y=k)=(1-p)^{k-1}p$$

There are k trials. Probability of success is p, probability of failure is $1-p$.

The probability that the first $k-1$ trials will result in failure:

$$(1-p)^{k-1}$$

What is the probability that the k^{th} trial results in success:

$$p$$

The probability of k trials, with the first $k-1$ resulting in failure, and the k^{th} resulting in success is given by:

$$P(Y=k)=(1-p)^{k-1}p$$

1.46: Mean of Geometric Distribution

$$E[Y]=\frac{1}{p}$$

1.47: Variance of Geometric Distribution

E. Uniform Distribution

1.48: Uniform Distribution

A random variable X with a uniform distribution, minimum value a, and maximum value b, has probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & if\ a \le x \le b \\ 0\ otherwise \end{cases}$$

- The uniform distribution has equal probability for all equal width intervals

1.49: Cumulative Probability Density Function

A random variable X with a uniform distribution, minimum value a, and maximum value b, has cumulative probability density function:

$$f(x) = \begin{cases} \frac{x-a}{b-a} & if\ a \le x \le b \\ 0\ otherwise \end{cases}$$

- The uniform distribution has equal probability for all equal width intervals

Example 1.50

Find $P(4 < X < 5)$ if X~U(4,10)

$$f(5) - f(4) = \frac{5}{10-4} - \frac{4}{10-4} = \frac{5}{6} - \frac{4}{6} = \frac{1}{6}$$

1.51: Mean of a Uniform Distribution

$$\mu = \frac{a+b}{2}$$

1.52: Variance of a Uniform Distribution

$$\sigma^2 = \frac{(b-a)^2}{12}$$

1.53: Cumulative Probability Density Function

A random variable X with a uniform distribution, minimum value a, and maximum value b, has cumulative probability density function:

$$f(x) = \begin{cases} \frac{x}{b-a} & if\ a \le x \le b \\ 0\ otherwise \end{cases}$$

The uniform distribution has equal probability for all equal width intervals

Example 1.54

You are going to meet your friend. He will meet you anytime between 1 to 4 pm (with equal probability for any specific time).

What is the probability that
A. you meet him between 1 to 4 pm?
B. you meet him between 1 to 2 pm?
C. you meet him at 2 pm?

Part A
1

Part B
1/3

Part C
0

F. Normal Distribution

The pdf of the Normal Distribution is mathematically complicated.

1.55: Normal Distribution

A random variable X with mean μ and variance σ^2 that follows a Normal Distribution is written

$$\underbrace{X}_{\substack{Random\\Variable}} \sim \underbrace{N(\mu,\sigma^2)}_{\substack{N=Normal\\Distribution}}$$

- The normal distribution is completely described (parametrized) by its mean and variance.

Example 1.56

X is a random variable that follows a Normal Distribution with mean 3 and variance 4. Y is a random variable that follows a Normal Distribution with mean 3 and standard deviation 2 and mean 3. Compare X and Y.

$$X \sim N(3,4)$$
$$Y \sim N(3,2^2) = Y \sim N(3,4)$$

Since the above are equal, they represent the same normal distribution.

1.57: Range of Values for X

A random variable X that follows a normal distribution can take any value on the real number line. That is

$$X\ can\ be\ (-\infty,\infty)$$

- However, note that the major proportion of values lies close to the mean.

Example 1.58

X is a random variable that follows a Normal Distribution with mean 3 and variance 4. The probability that X takes a value greater than 2023 is:

A. Zero
B. Non-zero
C. Cannot be determined
D. Negative

$$Non - zero$$

1.59: Probability for a specific value

Since the normal distribution is continuous, the probability for any specific value is always zero.

Example 1.60

Mark the correct option

X is a random variable that follows a Normal Distribution with mean 3 and variance 4. The probability that X takes the value 2023 is:

A. Zero
B. Non-zero
C. Cannot be determined
D. Negative

$$Zero$$

1.61: Symmetric Distribution

The normal distribution is symmetric.

- If you reflect it about its mean, the distribution remains unchanged.
- In a symmetric distribution $mean = median = mode$.

1.62: Mode of the Normal Distribution

The normal distribution is unimodal. That it has a single peak, which is its mode.

1.63: Bell Shaped

The normal distribution has a typical bell shape. Changing the

1.64: Area under the curve

The sum of the areas under the curve of the normal probability distribution is 1.

Example 1.65

A. X is a random variable that follows a Normal Distribution with mean μ and variance σ^2. Determine the probability that X is less than μ.
B. X is a random variable that follows a Normal Distribution with mean μ and variance σ^2. Determine the probability that X is less than or equal to μ.

Part A

Since the distribution is symmetric, the areas to the right and the left of the mean are equal, and hence the probabilities are also equal:

$$P(X < \mu) = \frac{1}{2}$$

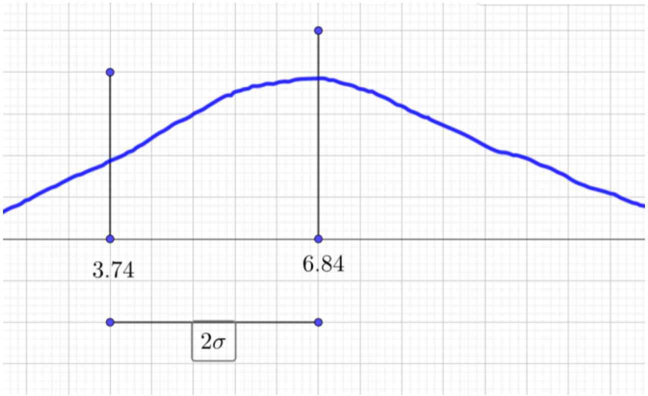
Part B

Since the probability of X taking a particular value in a continuous probability distribution is zero, the probability for $P(X \le \mu)$ is the same as $P(X < \mu)$:

$$P(X \le \mu) = P(X < \mu) = \frac{1}{2}$$

1.66: $68 - 95 - 99.7$ Rule

Example 1.67



Example 1.68

The probability that we want is:

$$P(\mu - \sigma < X < \mu + 2\sigma)$$

We can split the probability into two parts, and calculate each probability separately:

$$P(\mu - \sigma < X < \mu) + P(\mu < X < \mu + 2\sigma)$$

By symmetry,

$$P(\mu - \sigma < X < \mu) = \frac{68}{2} = 34\%$$
$$P(\mu < X < \mu + 2\sigma) = \frac{95}{2} = 47.5\%$$

And hence, the final answer is:

$$= 34 + 47.5 = 81.5\%$$

G. Standard Normal Distribution

Since the normal distribution is mathematically complicated to work with, we introduce the standard normal distribution. This distribution is a transformation of the normal distribution

1.69: Standard Normal Distribution

A random variable Z that follows a normal distribution with mean 0 and variance 1 is said to follow a standard normal distribution.

- The standard normal distribution is a special case of the normal distribution

1.70: Transformation from Normal to Standard Normal Distribution

If X follows a normal distribution with mean μ and variance σ^2, and Z that follows a normal distribution with mean 0 and variance 1, then you can convert from one to the other using the relation:

$$Z = \frac{X-\mu}{\sigma}$$

- Subtracting μ is a horizontal shift
- Dividing by σ is a vertical scale.

Example 1.71

We want the probability that:

$$P(X > 290)$$

Convert to a standard normal random variable:

$$P\left(\frac{X-\mu}{\sigma} > \frac{290-304}{8}\right)$$
$$P(Z > -1.75)$$
$$1 - P(Z < -1.75)$$
$$1 -$$

1.4 Sampling Distributions

A. Sampling Distributions

1.72: Parameters and Statistics

A parameter is a number that describes some characteristic of the population.
A statistic is a number that describes some characteristic of a sample.

Example 1.73

Identify the population of interest, parameter, the sample size and the statistic in each of the below:

> A. There are 50 lakh people in a Lok Sabha Constituency. Out of those 50 lakh people, 30 lakh people intend to vote for Candidate A in an election. A simple random sample conducted by a psephologist polls 300 people and determines that 200 of them say they vote for candidate A.

The population of interest is:

$$50\ Lakh\ people$$

The parameter is proportion of people who intend to vote for Candidate A

$$Population\ Proportion = p = \frac{Successful\ Outcomes}{Total\ Outcomes} = \frac{30\ Lakh}{50\ Lakh} = 60\%$$

The sample size is the number of people polled

$$= 300$$

The statistic is the proportion of people in the sample who say they intend to vote for candidate A:

$$Sample\ Proportion = \hat{p} = \frac{200}{300} = 66\frac{2}{3}\% = 66.\bar{6}\%$$

Example 1.74

I

1.75: Sampling Distribution

The sampling distribution of a statistic is the distribution of the values taken by the statistic in all possible samples of the same size from the same population.

- A sampling distribution is itself a probability distribution. (It is a special case of a probability distribution).

Example 1.76

An urn contains 9 green balls, and one yellow ball. You draw a random sample of two balls, without replacement. State the sampling distribution of the sample.

Outcome	GG	GY
$Probability$	$\frac{9}{10}\cdot\frac{8}{9}=\frac{8}{10}=\frac{4}{5}$	$\frac{9}{10}\times\frac{1}{9}=\frac{1}{10}$ $\frac{1}{10}\times\frac{9}{9}=\frac{1}{10}$ $Total=\frac{1}{5}$

$$\frac{4}{5}+\frac{1}{5}=\frac{5}{5}=1$$

B. Estimators

1.77: Unbiased Estimator

A statistic used to estimate a parameter is an unbiased estimator if the mean of its sampling distribution is equal to the value of the parameter being estimated.

C. Sample Proportions

1.78: Sampling Distribution of the sample proportion: $\hat{p}$

Choose a random sample size n from a large population of size N with proportion p of successes. Let the number of number of successes in the sample be the random variable X. Let the random variable $\hat{p}$ be the sample proportion of successes. That is:

$$\hat{p}=\frac{X}{n}$$

Assuming sampling with replacement, X is a random variable that follows:

$$X\sim Binomial(n,p), E[X]=np, Var[X]=np(1-p)$$

1.79: Mean of $\hat{p}$

The mean of the sample proportion is equal to the population proportion:

$$\mu_{\hat{p}}=p$$

The mean of the sample proportion is:

$$E[\hat{p}]=E\left[\frac{X}{n}\right]=\frac{1}{n}E[X]=\frac{1}{n}\cdot np=p$$

1.80: Standard Deviation of $\hat{p}$

The standard deviation of the sample proportion is given by:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \ if \ \underbrace{\frac{n}{N} \le 10\%}_{10\%\ Condition}$$

Assuming sampling with replacement, the variance of the sample proportion is:

$$Var[\hat{p}] = Var\left[\frac{X}{n}\right] = \frac{1}{n^2}Var[X] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$SD[\hat{p}] = \sqrt{Var[\hat{p}]} = \sqrt{\frac{p(1-p)}{n}}$$

If the sample size is small relative to the population size, then the above formula is a good approximation even for sampling without replacement. This is captured in the condition:

$$\frac{n}{N} \le 10\%$$

That is the sample size should be less than 10% of the population size.

1.81: Distribution of $\hat{p}$

If the 10% condition holds, then the distribution of $\hat{p}$ is approximately Binomial with parameters

$$(n, p)$$

The large counts condition is:

$$\underbrace{np \ge 10}_{Large\ Counts\ I} \quad , \quad \underbrace{n(1-p) \ge 10}_{Large\ Counts\ II}$$

If the large counts condition holds, then we can approximate the Binomial using a Normal Distribution

Example 1.82

Among a thousand monitors, there are exactly ten monitors with dead spots. A random sample of five monitors is checked for dead spots. In the sample, what is the:

A. distribution of the number of monitors that have a dead spot? (Since $1000 \gg 5$, assume sampling is with replacement.)
B. proportion of monitors that have dead spots?
C. Mean and standard deviation of the proportion of monitors that have dead spots?
D. can you approximate the distribution of the proportion of monitors that have a dead spot using a Normal Distribution?

Part A

Let X be the number of monitors that have dead spots.

$$X \sim Binomial(5, 0.01)$$

$$P(X = x) = \binom{5}{x}(0.01)^5(0.99)^{5-x}$$

Part B

$$\hat{p} = \frac{X}{n} = \frac{X}{5}$$

Part C

The mean of the sample proportion is equal to the population proportion:

$$\mu_{\hat{p}} = p = 0.01$$

$$\frac{n}{N}=\frac{5}{1000}=0.5\%<10\%$$

$$\sigma_{\hat{p}}=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.01(0.99)}{5}}$$

Part D

$$np=5(0.01)=0.05$$

D. Sample Means

1.83: Mean of $\bar{x}$

A simple random sample of size n is drawn from a large population of size N with mean μ and standard deviation σ.

The sample mean $\bar{x}$ has a sampling distribution with:

$$Mean=\mu_{\bar{x}}=\mu$$

- Sample mean is an unbiased estimator of the population mean.

$$E[\bar{x}]=E\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]$$
$$=\frac{1}{n}(E[X_1]+E[X_2]+\cdots+E[X_n])$$
$$=\frac{1}{n}(\mu+\mu+\cdots+\mu)$$
$$=\frac{1}{n}(n\mu)$$
$$=\mu$$

1.84: Standard Deviation of $\bar{x}$

A simple random sample of size n is drawn from a large population of size N with mean μ and standard deviation σ.

The sample mean $\bar{x}$ has a sampling distribution with:

$$Standard\ Deviation=\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}\ if\ \ \underbrace{\frac{n}{N}\le 10\%}_{10\%\ Condition}$$

The variance is given by:

$$Var[\bar{x}]=Var\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]$$
$$=\frac{1}{n^2}(Var[X_1]+Var[X_2]+\cdots+Var[X_n])$$
$$=\frac{1}{n^2}(\sigma^2+\sigma^2+\cdots+\sigma^2)$$
$$=\frac{1}{n^2}(n\sigma^2)$$
$$=\frac{\sigma^2}{n}$$

$$SD[\bar{x}] = \sqrt{Var[\bar{x}]} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

We know the mean and the standard deviation of the sample mean. But we do not know the distribution that it follows. We would ideally like the distribution to be normal. There are two cases in which this will happen:

- The population is normal.
- The sample size is > 30.

1.85: Sampling from a Normal Population

If you draw a simple random sample from a normal population with mean μ and standard deviation σ, the sample mean follows a normal distribution.

1.86: Sampling from a Normal Population

If you draw a simple random sample of size $n > 30$ from a population with mean μ and standard deviation σ, the sample mean follows a normal distribution.

Example 1.87

Concept of Sampling Distribution

You get a shipment of 1000 bulbs of brand X. Brand X promises that its lightbulbs last 300 hours (on average). You select 5 bulbs for a test check. The mean lifetime of those 5 lightbulbs is 203 hours.

Population parameter = mu = 300 hours
Sample statistic = x bar = 203 hours

Sampling Distribution
The number of possible samples of size 5 that you can take from 1000 bulbs is
1000C5=Y

If you calculate the mean of each of the Y possible samples, and then plot that distribution, that is the sampling distribution of the sample mean.

Mean of the sampling distribution
=Mean of population
=300

E. Sample Proportion

1.88: Sample Proportion and Population Proportion

Percentage of the sample that is “successful” is sample proportion.
Percentage of the population that is “successful” is population proportion.

1.89: Sample Proportion

The sample mean is an unbiased estimator of the population.

$$E(\bar{X}) = P$$

1.90: Standard Error of the Proportion

Standard error of the mean is

$$\sigma_{\bar{X}} = \sqrt{\frac{P(1-P)}{n}}$$

1.91: Converting to Z

$$Z = \frac{\bar{X} - P}{\sigma_{\bar{X}}} = \frac{\bar{X} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

1.5 Confidence Intervals

A. Estimation and Inference

1.92: Inference

Inference is making statements about the population parameters using the sample statistic.

1.93: Estimator

The sample statistic is used to “estimate” the value of the population parameter.

- The point estimator equals the sample statistic
- Important: Is the sample statistic a good estimator of the population parameter?

1.94: Unbiased Estimator

The sample statistic is an unbiased estimator of the population parameter if the mean of the sampling distribution equals the population parameter.

$$E(\hat{\theta}) = \theta$$

1.95: Bias

Bias of an unbiased estimator is zero.

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

1.96: Efficient Estimator

The most efficient estimator of θ

- Is one where the estimator is an unbiased estimator
- with the smallest variance.

B. Confidence Intervals for Mean

1.97: Confidence Interval for the Mean

$$\sigma\ known: \bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\sigma\ unknown: \bar{X} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

- Example: We are 95% confident that the true mean lies within the interval.

We use Z table when population standard deviation $= \sigma$ is known.
We use t table when population standard deviation $= \sigma$ is unknown.

$$t = \frac{X - \mu}{s/\sqrt{n}}$$

t distribution has wider confidence intervals since the population standard deviation is not known.
As $n - 1 = df$ increases, the t distribution becomes closer to the Z distribution.

Example 1.98

Find $Z_{\frac{\alpha}{2}}$ for a 95% confidence interval

Probability to find Z:

$$95 + \frac{100 - 95}{2} = 95 + \frac{5}{2} = 95 + 2.5 = 97.5\%$$

Probability to find t:

$$95 + \frac{100 - 95}{2} = 95 + \frac{5}{2} = 95 + 2.5 = 97.5\%$$

1.99: Margin of Error

$$ME = Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced (σ↓)
- The sample size is increased (n↑)
- The confidence level is decreased, (1 – a) ↓

C. Confidence Intervals for Proportions

1.100: Confidence Interval for Proportions

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

1.6 More Estimation

A. Dependent and Independent Samples

1.101: Dependent Samples

A sample before/after a certain condition.

Average score before joining a preschool program

$$= E(X)$$

Average score after joining a preschool program

$$= E(Y)$$

Difference of average scores

$$= E(Y) - E(X)$$

Dependent samples are taken at different time points on the same population

1.102: Independent Samples

Independent samples are samples taken from two different populations

B. Mean: Dependent Samples

1.103: Dependent Samples: Difference in Sample Means

$$\bar{d} = \frac{\sum d_i}{n} = \frac{\sum (y_i - x_i)}{n}$$

$$y_i = score\ of\ i^{th}\ child\ (After)$$
$$x_i = score\ of\ i^{th}\ child\ (Before)$$

1.104: Dependent Samples: Confidence Intervals:

$$\sigma\ known: \bar{d} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma_d}{\sqrt{n}}$$

C. Mean: Independent Samples

1.105: Independent Samples: Difference in Means:

We are looking at two different sample groups.

1.106: Independent Samples: Difference in Means

Difference between means when sample are drawn from populations X and Y, with populations means μ_x and μ_y:

$$Point\ Estimate: \bar{x} - \bar{y}$$
$$\bar{x} = Mean\ of\ sample\ from\ population\ X$$
$$\bar{y} = Mean\ of\ sample\ from\ population\ Y$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$
$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum y_i}{n}$$

The samples are not paired, so we cannot use the same formula as we used for dependent.

For example,

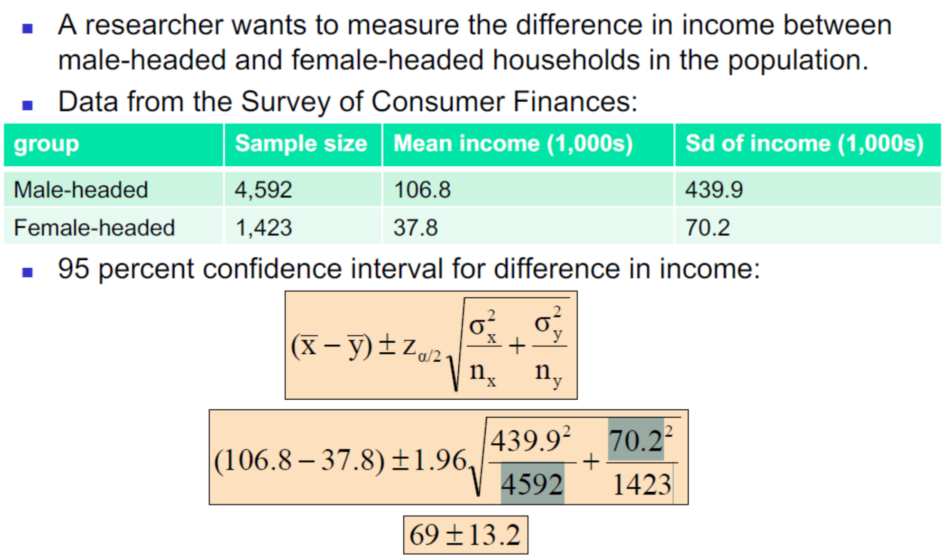
- Do men get higher salaries compared to women?

1.107: Independent Samples: Difference in Means: Standard Error

$$\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

1.108 Independent Samples: Difference of Means: Confidence Interval

$$(\bar{x}-\bar{y}) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$



- A researcher wants to measure the difference in income between male-headed and female-headed households in the population.
- Data from the Survey of Consumer Finances:

group	Sample size	Mean income (1,000s)	Sd of income (1,000s)
Male-headed	4,592	106.8	439.9
Female-headed	1,423	37.8	70.2

- 95 percent confidence interval for difference in income:

$$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

$$(106.8-37.8) \pm 1.96\sqrt{\frac{439.9^2}{4592}+\frac{70.2^2}{1423}}$$

$$69 \pm 13.2$$

D. Proportions

1.109: Difference in Proportions: Point Estimate

This is the difference in the sample proportions:

$$\hat{p}_x-\hat{p}_y$$

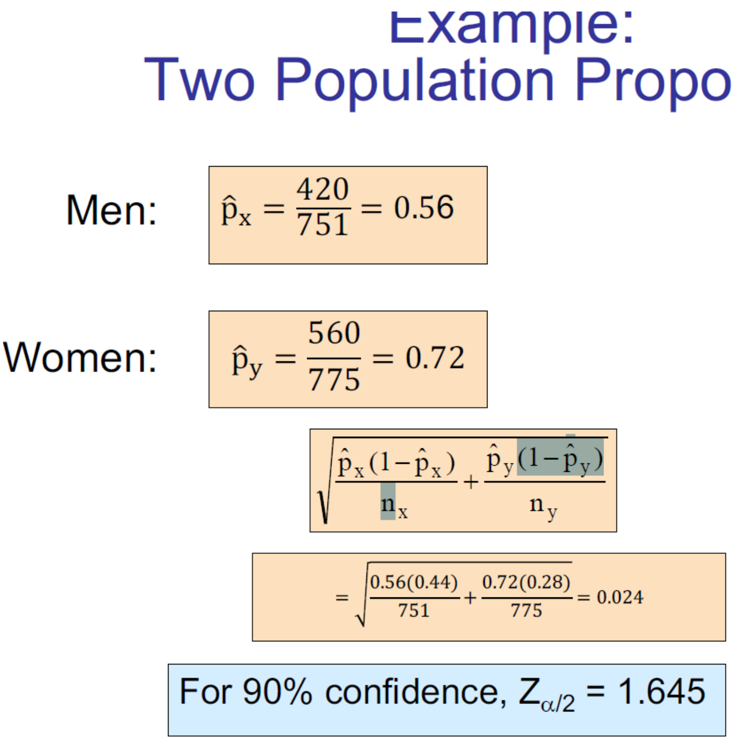
1.110: Difference in Proportions: Standard Error

$$\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

1.111: Difference in Proportions: Confidence Intervals

$$(\hat{p}_x - \hat{p}_y) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

How confident are we that the difference in proportions is this.



Example:
Two Population Proportion

Men: $\hat{p}_x = \frac{420}{751} = 0.56$

Women: $\hat{p}_y = \frac{560}{775} = 0.72$

$$\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

$$= \sqrt{\frac{0.56(0.44)}{751} + \frac{0.72(0.28)}{775}} = 0.024$$

For 90% confidence, $Z_{\alpha/2} = 1.645$

1.7 Hypothesis Testing

A. Basics

1.112: Testing a Claim

Research/statistics are used to test claims about any number of things. Research is useful in:

- Medicine: Testing a new drug
- Education: Comparing two groups of children to see which performs better
- Sociology: Comparing two populations to see their cultural attributes

1.113: Null and Alternate Hypothesis

The null and alternate hypothesis are always about a population parameter, not a sample statistic.

- We start research by formulating a null and an alternate hypothesis.
- These need to be framed before the data collection.

1.114: One-Tailed Versus Two Tailed Test

One tailed test is when you have one interpretation. Eg: It increases, etc

One tailed test is when you have two interpretations. Eg: It increases or decreases.

1.115: Confidence Level

The level of confidence that you wish to test the hypothesis at is:

$$1-\alpha$$

- Changing the level of confidence will the numerical calculations for a statistical question.

B. Testing for a Single Mean

1.116: Confidence Interval Method

In the confidence interval method, you use the sample statistic to generate a confidence interval related to the population parameter. If the population parameter falls outside the confidence interval, you reject the null hypothesis.

Example 1.117

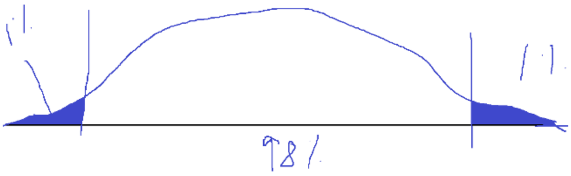
The mean value of hemoglobin in a sample of 100 children from a district is 9.1%. If the hemoglobin level for healthy children is 13% , and the population standard deviation of hemoglobin is 2%. Determine, at a 98% level of confidence, whether children in the district have hemoglobin different from that for healthy children.

This is a one tailed test
You start with the opposite hypothesis of what you wish to show is true.

$$Null\ Hypothesis: H_0: u_0 = 13\%$$

$$Alternate\ Hypothesis: H_A: u_A < 13\%$$

$$\bar{X} = Mean\ Hemoglobin\ in\ blood\ samples = 9.1\%$$



Find Z at 98% level of confidence:

$$Z_{\frac{0.02}{2}} = 2.34$$

Find the confidence interval:

$$\bar{X} \pm Z_{\frac{0.02}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 9.1\% \pm 2.34 \cdot \frac{2\%}{\sqrt{100}} = 9.1\% \pm 0.468 \Rightarrow [8.632, 9.568]$$

$u_0 = 13\%$ does not fall in the confidence interval.
Hence, we reject the null hypothesis.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{9.1\% - 13\%}{2\%/\sqrt{100}} = \frac{-3.9\%}{0.2\%} = -19.5$$

Example 1.118

Normal blood pressure is 120. Give blood pressure medication to a group of patients. Current average blood pressure of the patients is 150. You want evidence to submit to the FDA that the medication works. Write the null

$$\mu_x = Average\ blood\ pressure\ at\ start\ of\ trial$$
$$\mu_y = Average\ blood\ pressure\ after\ six\ weeks$$

We are interested in the population parameter:

$$\mu_y - \mu_x$$

Null Hypothesis: There is no change: $\mu_y - \mu_x = 0$
Alternate Hypothesis: $\mu_y - \mu_x < 00$

Consider a sample of 100 patients.

$$n = 100$$

Calculate the sample statistic.

$$\bar{d} = sample\ difference\ in\ means\ (before\ and\ after\ taking\ the\ medication\ for\ six\ weeks)$$

Use that $\bar{d}$ to find a 95% confidence interval $\mu_y - \mu_x$ using dependent samples. Note that this is a one-tailed test.

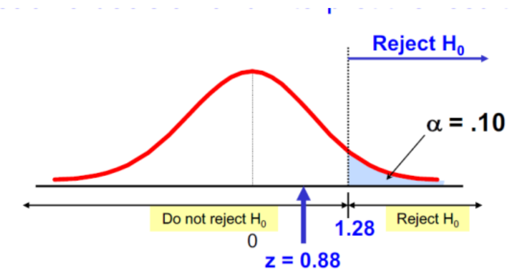
If Null Hypothesis value $(\mu_y - \mu_x = 0)$ falls outside the confidence interval, then and we reject the Null Hypothesis at the 5% level of confidence.

1.119: Critical value approach

The critical value is the value at which the null hypothesis will be rejected.

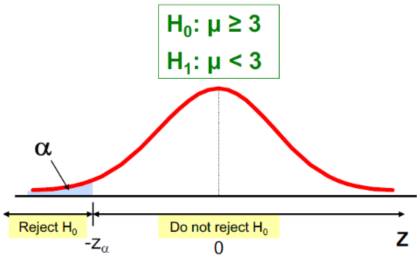
1.120: Upper Tail Test

$$H_0: \mu \le A, \qquad H_A: \mu > A$$
$$Z < Z_\alpha \Rightarrow Fail\ to\ reject\ null\ hypothesis$$
$$Z > Z_\alpha \Rightarrow Reject\ null\ hypothesis$$



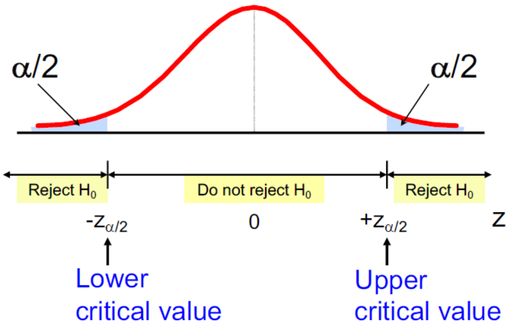
1.121: Lower Tail Test

$$H_0: \mu \ge A, \qquad H_A: \mu < A$$
$$Z > -Z_\alpha \Rightarrow Fail\ to\ reject\ null\ hypothesis$$
$$Z < -Z_\alpha \Rightarrow Reject\ null\ hypothesis$$



1.122: Two Tailed Test

$$H_0: \mu = A, \qquad H_A: \mu \ne A$$
$$-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}} \Rightarrow Fail\ to\ reject\ null\ hypothesis$$
$$Z > -Z_{\frac{\alpha}{2}}\ OR\ Z < -Z_{\frac{\alpha}{2}} \Rightarrow Reject\ null\ hypothesis$$

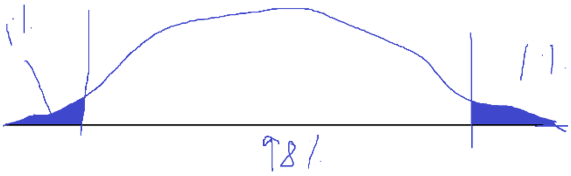


Example 1.123

The mean value of hemoglobin in a sample of 100 children from a district is 9.1%. If the hemoglobin level for healthy children is 13% , and the population standard deviation of hemoglobin is 2%. Determine, at a 98% level of confidence, whether children in the district have hemoglobin different from that for healthy children.

This is a one tailed test
You start with the opposite hypothesis of what you wish to show is true.

$$Null\ Hypothesis: H_0: u_0 = 13\%$$
$$Alternate\ Hypothesis: H_A: u_A \neq 13\%$$



$$\bar{X} = Mean\ Hemoglobin\ in\ blood\ samples = 9.1\%$$

Find critical value of Z at 98% level of confidence:

$$-Z_{\frac{0.02}{2}} = -2.34$$

Hence, we reject the null hypothesis.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{9.1\% - 13\%}{2\%/\sqrt{100}} = \frac{-3.9\%}{0.2\%} = -19.5$$

Since

$$-19.5 < -2.34$$

We reject the null hypothesis at the 98% level of confidence.

Example 1.124

To test the claim that mean cell phone bills are greater than \$52 at $\alpha = 0.1$, a sample size of 64 phone bills was taken and the sample mean was determined to 53.1. A prior study informs you that the population standard deviation for cell phone bills is \$10.

Given data:

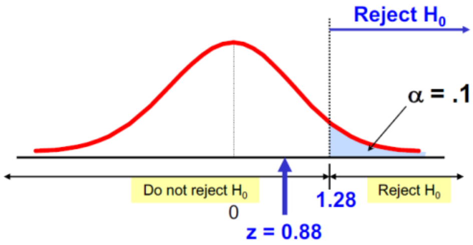
$$\sigma = 10, \qquad n = 64, \qquad \bar{X} = 53.1$$

Decide your null and alternate hypothesis:

$$H_0: \mu_0 \leq 52, \qquad H_A: \mu_0 > 52$$

Find the level of confidence:

$$\alpha = 0.1 \Rightarrow Level\ of\ Confidence = 1 - \alpha = 1 - 0.1 = 0.9 = 90\%$$



Find the critical Z value:

$$Z_\alpha = Z_{0.1} = 1.28$$

Decision Criterion:

$$Reject\ H_0\ if\ Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.28$$

Calculate Z

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{53.1 - 52}{10/\sqrt{64}} = \frac{1.1}{10/8} = 0.88$$

We fail to reject the null hypothesis.

1.125: *p* value approach

The p value is the probability of getting a sample statistic, or worse, if the null hypothesis is true.

Decision Criterion:

$$p\ value > \alpha \Rightarrow Fail\ to\ reject\ null\ hypothesis$$
$$p\ value < \alpha \Rightarrow Reject\ null\ hypothesis$$

Example 1.126

To test the claim that mean cell phone bills are greater than \$52 at $\alpha = 0.1$, a sample size of 64 phone bills was taken and the sample mean was determined to 53.1. A prior study informs you that the population standard deviation for cell phone bills is \$10.

$$\sigma = 10, \quad n = 64, \quad \bar{X} = 53.1, \quad H_0: \mu \le 52, \quad H_A: \mu > 52$$

Calculate the p value:

$$P(\bar{X} > 53.1 \mid \mu = 52) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{53.1 - 52}{10/\sqrt{64}}\right) = P(Z > 0.88) = 0.18943$$

Compare with α

$$p\ value = 0.18943 > 0.1 \Rightarrow Fail\ to\ reject\ the\ null\ hypothesis$$

C. Testing for Proportions

1.127: Testing for Proportion

The test statistic for sample proportion follows a Z distribution:

$$Z = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

D. Type I and Type II Error

1.128: Type I Error

Rejecting a true null hypothesis.

$$Probability\ of\ Type\ I\ Error = \alpha$$

- This is also called the level of significance.
- It is supposed to be set in advance of your test.

1.129: Type II Error

Failing to reject a false null hypothesis.

$$Probability\ of\ Type\ I\ Error = \beta$$

- Type I and Type II errors cannot happen at the same type.

Example 1.130

If a defendant is put on trial for a murder. The presumption in law is “innocent until proven guilty”. Identify Type I and Type II error in this context.

Null Hypothesis: The defendant is innocent.
Alternate Hypothesis: The defendant is guilty.

	$H_0\ is\ true: Innocent$	$H_A\ is\ true: Guilty$
Fail to reject H_0	$Correct$ $1-\alpha$	$Type\ II\ Error$ β
Reject H_0	$Type\ I\ Error$ α	$Correct$ $Power: 1-\beta$

1.131: Significance Level and Power

The significance level of a test is the probability of reaching the wrong conclusion when the null hypothesis is true.

$$Significance\ Level = P(Type\ I\ Error) = \alpha$$

1.132: Significance Level and Power

The power of a test to detect a specific alternative is the probability of reaching the right conclusion when that alternative is true.

$$Power = 1-\beta$$

E. Extras to Remember

1.133: Standard Deviation

The standard deviation is the square root of the average squared deviation from the mean.

$$Sample\ Standard\ Deviation = s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$Population\ Standard\ Deviation = \sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

F. Difference of Means

1.134: Difference of Means: Dependent Samples

For the null hypothesis where the means are assumed to be equal, the test statistic is:

$$Z = \frac{\bar{d} - 0}{\sigma_d / \sqrt{n}}$$

1.135: Difference of Means: Independent Samples

For the null hypothesis where the means are assumed to be equal, the test statistic is:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

G. Difference of Proportions

1.136: Difference of Proportions: Independent Samples:

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - 0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}$$

$$Pooled\ Proportion = \hat{p}_0 = \frac{n_x}{n_x + n_y}\hat{p}_x + \frac{n_y}{n_x + n_y}\hat{p}_y$$

1.8 Inference for Categorial Data

A. Chi Square Tests for Goodness of Fit

1.137: Goodness of Fit

Goodness of fit determines whether a data set fits a given distribution.

1.138: χ^2 Test Statistic

The χ^2 Test Statistic measures the difference between the observed data and the actual data.

$$\chi^2 = \sum \frac{(O - E)^2}{E^2}$$

Where

$$O = Observed$$
$$E = Expected$$

1.139: χ^2 Distribution

The χ^2 Test Statistic follows a χ^2 distribution with $n - 1$ degrees of freedom, where n is the number of categories.

B. Inference for Two Way Tables

1.9 Regression

A. Inference for Regressions

1.140: Conditions for linear inference

- Linear: The actual relationship between x and y is linear. For any fixed value of x, the mean response my falls on the population (true) regression line $\mu_y = \alpha + \beta x$.
- Independent: Individual observations are independent of each other. When sampling without replacement, check the 10% condition.
- Normal: For any fixed value of x, the response y varies according to a Normal distribution.
- Equal SD: The standard deviation of y (call it \sigma) is the same for all values of x.
- Random: The data come from a well-designed random sample or randomized experiment.

1.141: Regression Line

$$Population\ Regression\ Line: y = \alpha + \beta x$$
$$\alpha = Population\ Intercept$$
$$\beta = Population\ Slope$$

$$Sample\ Regression\ Line: \hat{y} = a + bx$$

1.142: Sampling Distribution of *b*

$$Mean = \mu_b = \beta$$
$$Standard\ deviation = \sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$$

$$\sigma = Standard\ deviation\ of\ residuals\ for\ population\ regression\ line$$
$$\sigma_x = standard\ deviation\ of\ the\ explanatory\ variable\ x$$
$$n = sample\ size$$

1.143: Standardizing

$$Z = \frac{b - \beta}{\sigma_b}$$

1.144: s as estimator for σ

$$s = \sqrt{\frac{\sum Residuals^2}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}}$$

$$SE_b = \frac{s}{s_x\sqrt{n-1}}$$

1.145: Standardizing

$$t_{n-2\ df} = \frac{b - \beta}{SE_b}$$

1.146: Confidence Interval

$$b \pm t^* \cdot SE_b$$

1.147: Significance Test

$$Null\ Hypothesis: \beta = \beta_0$$

$$Test\ Statistic: t_{n-2\ df} = \frac{b - \beta_0}{SE_b}$$

Most common null hypothesis is

$$\beta = 0$$

B. Transformations

1.148: Exponential Model

$$y = ab^x$$

$$\ln y = \ln a + x \ln x$$

149 Examples