Introduction to Data Science

Unit 3 – Sampling and Estimation

Preet Kanwal
Assistant Professor
Department of CSE
PESU Bangalore

Unit 3 Contents

Outline - Chapter 4

(Listed in the order followed in class)

- Probability Plots (Section 4.10 – Page 285)
- The Central Limit Theorem (Section 4.11 – Page 290)
- Some principles of Point Estimation (Section 4.9 – Page 280)

Outline – Chapter 5

(Listed in the order followed in class)

Construction of Confidence Intervals for:

a) Population Mean -

Population mean of Large samples. (Section 5.1)

Population mean of Small samples. (Section 5.3)

Difference between Two Population means of Large samples. (Section 5.4)

b) Confidence Intervals with Paired Data (Section 5.7)

c) Population Proportion -

Proportions of Large Samples. (Section 5.2)

Difference between Two proportions of Large Samples. (Section 5.5)

Normal Probability plot

Prof. Preet Kanwal

How can we say our data came from normal population?

For Large Samples:
Shape of the histogram is symmetric.
No outliers.

Hence, Normal distribution should not be used for data sets that contain outliers.

Probability plots(or Quantile-Quantile plots) can be used to determine whether a reasonably large sample came from normal population.

For small samples:
Its difficult to determine – must have some knowledge about the process that generated the data.

Quantile

Quantile - splitting the distribution of a variable into n equally sized pieces.

The n quantiles for a data set are found approximately by ranking the data in order and then splitting this ranking through n - 1 equally spaced points on the interval.

If we have a probability density function for a continuous random variable, we use the above integral to find the quantiles. For n quantiles, we want:

1. The first to have 1/n of the area of the distribution to the left of it.
2. The second to have 2/n of the area of the distribution to the left of it.
3. The rth to have r/n of the area of the distribution to the left of it.
4. The last to have (n - 1)/n of the area of the distribution to the left of it.

Q-Q Plot

The quantile-quantile (q-q) plot is:

1. A graphical technique for determining if two data sets come from populations with a common distribution.
2. A q-q plot is a plot of sorted quantiles of the first data set against the sorted quantiles of the second data set. Q–Q plot follows the 45° line y = x. Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other

Advantages of the q-q plot are:

1) The sample sizes do not need to be equal.

2) Many distributional aspects can be simultaneously tested. For example:
if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

Probablity plot – important version of Q-Q plot

Here, 2nd distribution is a theoretical one (rather than that of a second data set) , we use the quantiles of a theoretical distribution and one is checking to see if the shape of the data set matches the theoretical shape.

The probability plot is:
1) A graphical technique for assessing whether or not a data set follows a given distribution such as the normal or Weibull.

2)The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line. Departures from this straight line indicate departures from the specified distribution.

3) The **normal probability plot** is used to answer the following questions.

a) Are the data normally distributed?

b) What is the nature of the departure from normality (data skewed, shorter than expected tails,

longer than expected tails)?

Constructing a Probability Plot

1) Sort the data.

2)Assign evenly spaced values to the data between 0 and 1.

For each x_i in the data set,

$$(i - 0.5) / n$$

Where,

i is the position of the data item

n is the size of the data set.

3) Find theoritical quantiles - Q_i.

4) Plot every point (x_i , Q_i).

5) Plot (x_i , x_i)

6) Observe visually whether the points form approximately a straight line. Departures from this straight line indicate departures from the specified distribution.

Other ways of dividing the data set equally between 0 and 1

Methods	Plotting Position $method(i,n)$
Blom	$(i-0.375)/(n+0.25)$
Benard	$(i-0.3)/(n+0.4)$
Hazen	$(i-0.5)/n$
Van der Waerden	$i/(n+1)$
Kaplan-Meier	i/n

Problem 1

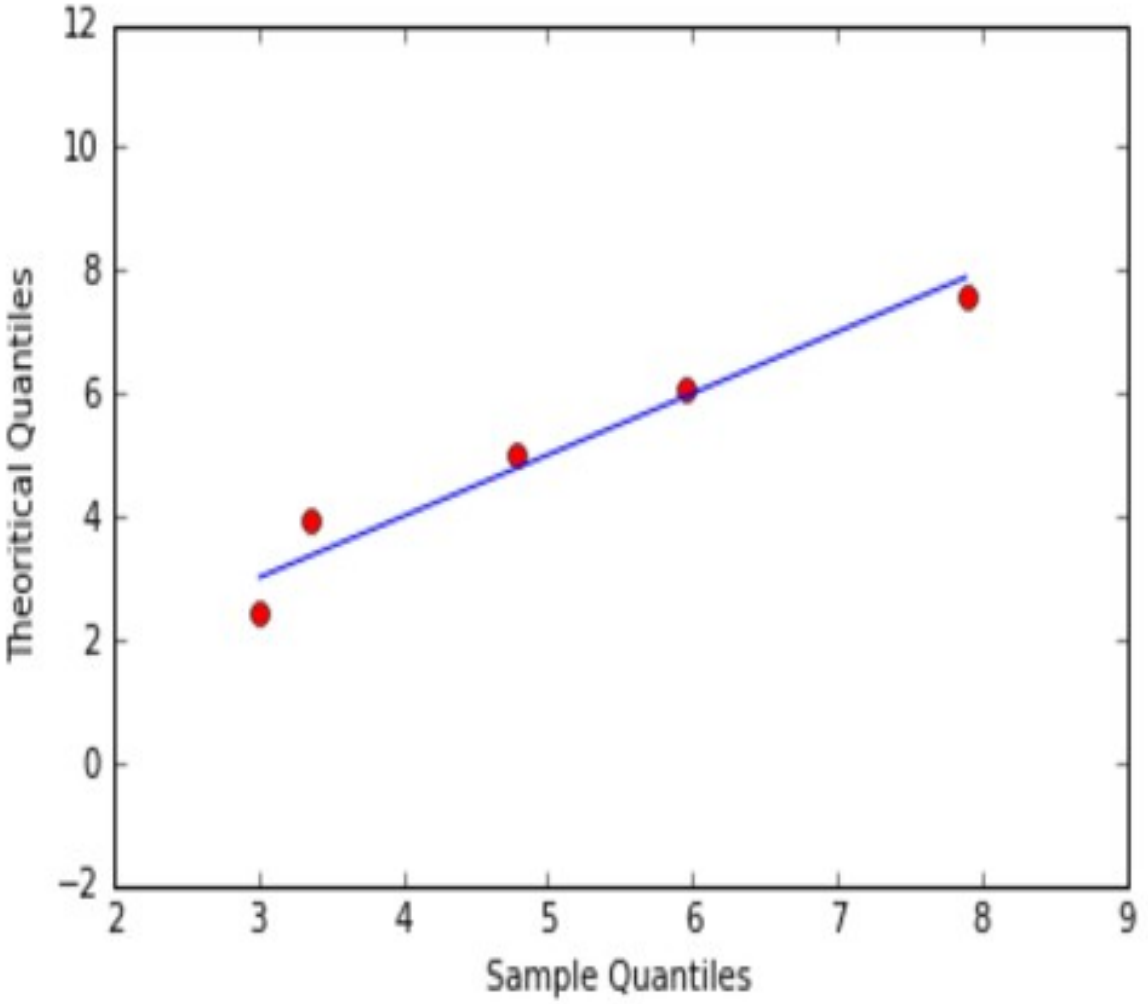
Construct a normal probability plot for the following data. Do these data appear to come from an approximately normal distribution?

3.01, 3.35, 4.79, 5.96, 7.89

Problem 1 - Solution

Positon (i)	Data items(X_i)	$(i - 0.5)/n$	Closest area in z-table	z-score	X value wrt to the z-score (Q_i) $X = z * sigma + mu$
1	3.01	(1 – 0.5)/5 = 0.1	0.1003	-1.28	-1.28 * 2 + 5 = 2.44
2	3.35	(2 – 0.5)/5 = 0.3	0.3015	-0.52	-0.52 * 2 + 5 = 3.95
3	4.79	(3 – 0.5)/5 = 0.5	0.5000	0.00	0 * 2 + 5 = 5.00
4	5.96	(4 – 0.5)/5 = 0.7	0.6985	0.52	0.52 * 2 + 5 = 6.05
5	7.89	(5 – 0.5)/5 = 0.9	0.8997	1.28	1.28 * 2 + 5 = 7.56

Problem 1 - Solution



Problem 2

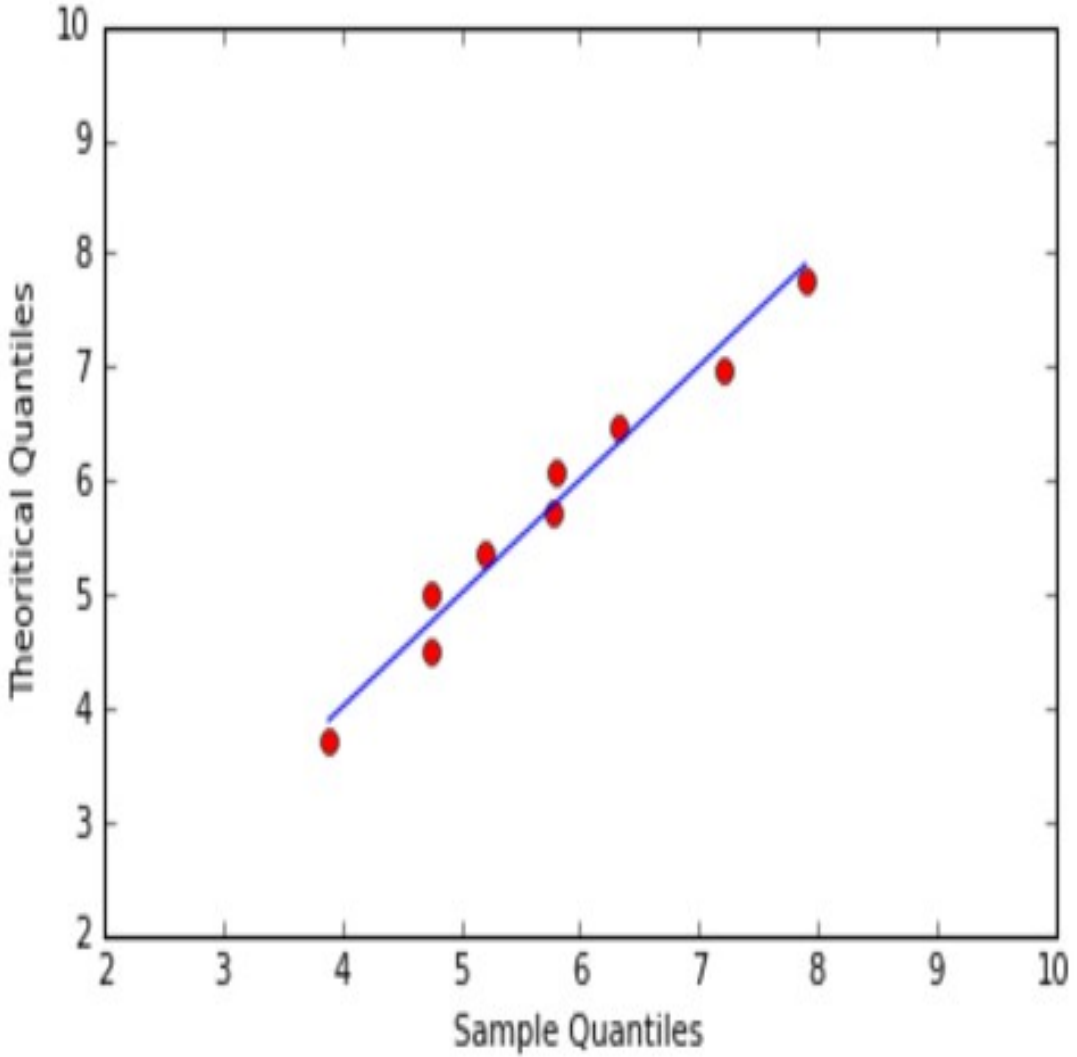
Construct a normal probability plot for the following data. Do these data appear to come from an approximately normal distribution?

3.89, 4.75, 4.75, 5.20, 5.78, 5.80, 6.33, 7.21, 7.90

Problem 2 - Solution

Positon (i)	Data items(X_i)	$(i - 0.5)/n$	Closest area in z-table	z-score	X value wrt to the z-score (Q_i) X = z * sigma + mu
1	3.89	(1 – 0.5)/9 = 0.0556	0.0559	-1.59	-1.59 * 1.268 + 5.734 = 3.71
2	4.75	(2 – 0.5)/9 = 0.1667	0.1660	-0.97	-0.97 * 1.268 + 5.734 = 4.5
3	4.75	(3 – 0.5)/9 = 0.2778	0.2776	-0.59	-0.59 * 1.268 + 5.734 = 4.98
4	5.20	(4 – 0.5)/9= 0.3889	0.3897	-0.28	-0.28 * 1.268 + 5.734 = 5.37
5	5.78	(5 – 0.5)/9 = 0.5000	0.5000	0.00	0 * 1.268 + 5.734 = 5.73
6	5.80	(6 – 0.5)/9 = 0.6111	0.6130	0.28	0.28 * 1.268 + 5.734 = 6.09
7	6.33	(7 – 0.5)/9 = 0.7222	0.7224	0.59	0.59 * 1.268 + 5.734 = 6.48
8	7.21	(8 – 0.5)/9 = 0.8333	0.8340	0.97	0.97 * 1.268 + 5.734 = 6.96
9	7.90	(9 – 0.5)/9 = 0.9444	0.9441	1.59	1.59 * 1.268 + 5.734 = 7.75

Problem 2 - Solution



Continuity Correction

Prof. Preet Kanwal

Continuity Correction

Its an adjustment made when approximating a discrete distribution with a contiuous one, so as to improve the accuracy of the approximation.

Here we'll discuss:
1) Normal approximation to Binomial

2) Normal approximation to Poisson

1) Normal approximation to Binomial

Normal Approximation to the Binomial

If $X \sim Bin(n, p)$ and,

If $np > 10$ and $n(1 - p) > 10$ (n is quite large) then,

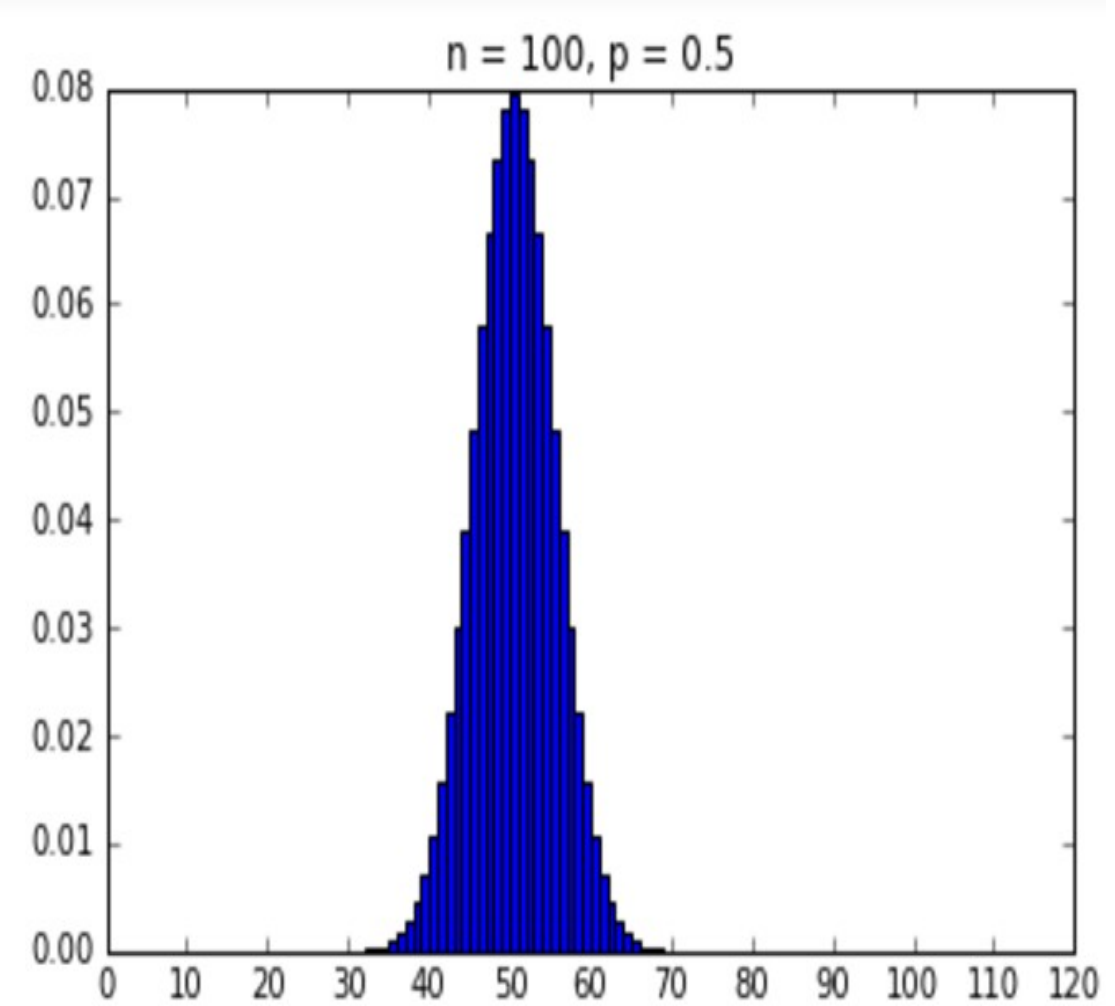
$$X \sim N(np , np(1 - p))$$ approximately.

Example : Bin(100, 0.5)

Suppose we wish to find P(45 <= X <=55).

The exact probability is given by total area of the rectangles of the binomial probability histogram corresponding to the integers 45 to 55 inclusive.

P(45 <= X <=55) = 0.7287

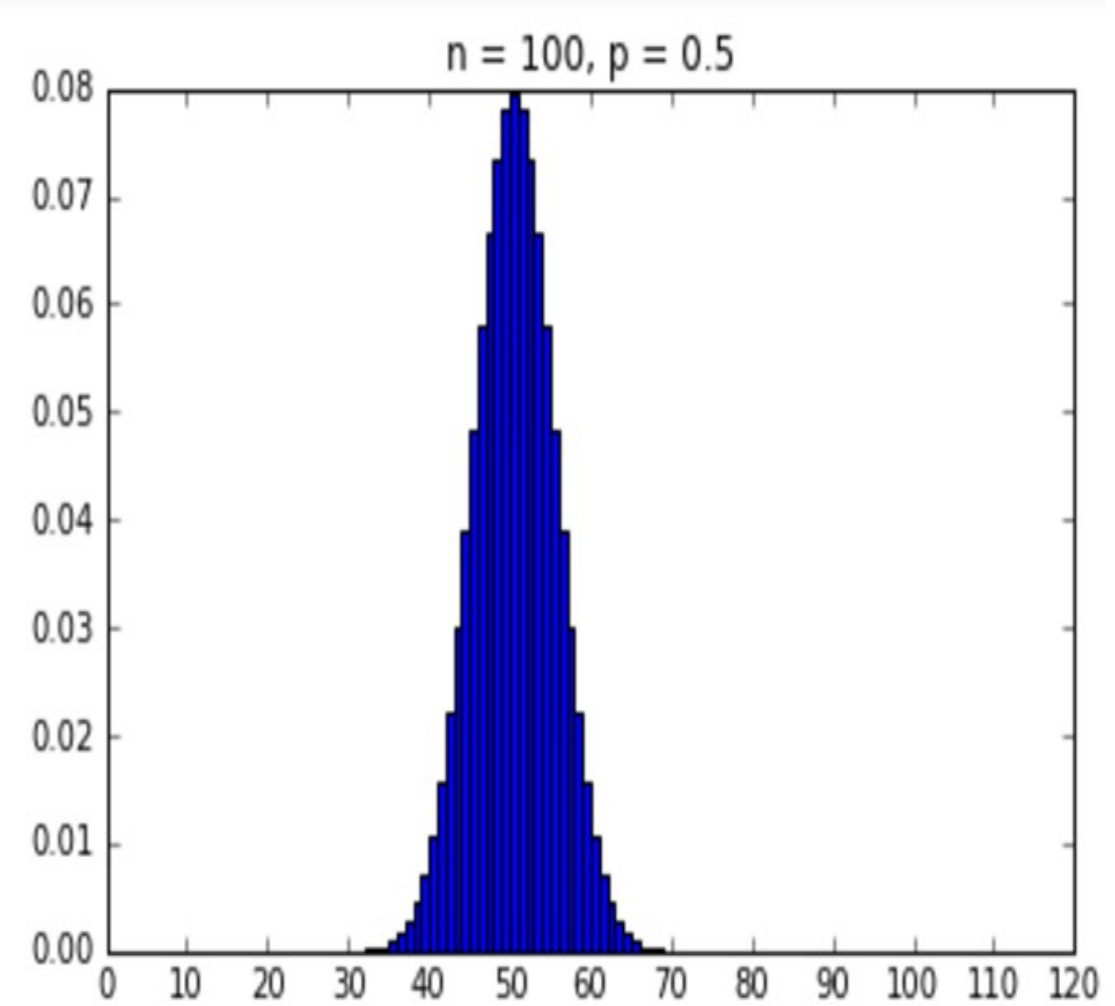


Bin(100, 0.5) can be approximated as $N(50, 5^2)$

We can use z-table to find P(45 <= X <=55).

But that would exclude the end points 45 and 55.
P(45 <= X <=55). = 0.6827

Incorporating Continuity Correction:
P(44.5 <= X <= 55.5) = 0.7287

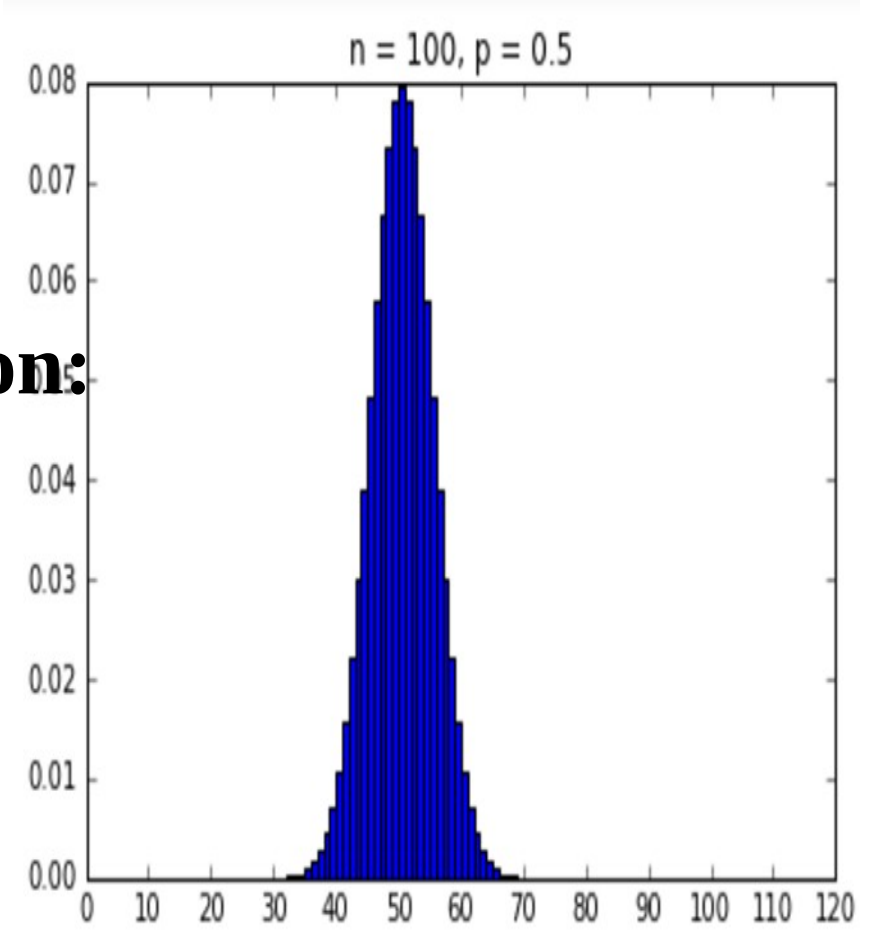


Computing probability that corresponds to an area in the tail of the distribution

Actual Probability
P(X >= 60) = 0.02844

Using z-table:
Without Continuity Correction:
Z = (60 - 50)/5 = 2
P(Z >=2) = 0.02275

Incorporating Continuity Correction:
P(X >= 55.5) = 0.1357



Accuracy of Continuity Correction

- It improves the accuracy of the normal approximation to binomial distribution in most cases.
- Can reduce the accuracy of normal approximation to binomial distribution when computing probbaility that corresponds to an area in the tail of the distribution.
- This results from the fact that the normal approximation is not perfect. It fails to account for a small degree of skewness in the distribution.

2) Normal approximation to Poisson

Normal Approximation to the Poisson

If $X \sim \text{Poisson}(\lambda)$ where $\lambda > 10$, then

$$X \sim N(\lambda , \lambda) \text{ approximately.}$$

Continuity Correction for Poisson distribution

- For areas that include the central part of the curve, the continuity correction generally imporves the normal approximation.

- But, for areas in the tails, the continuity correction sometimes makes the approximation worse.

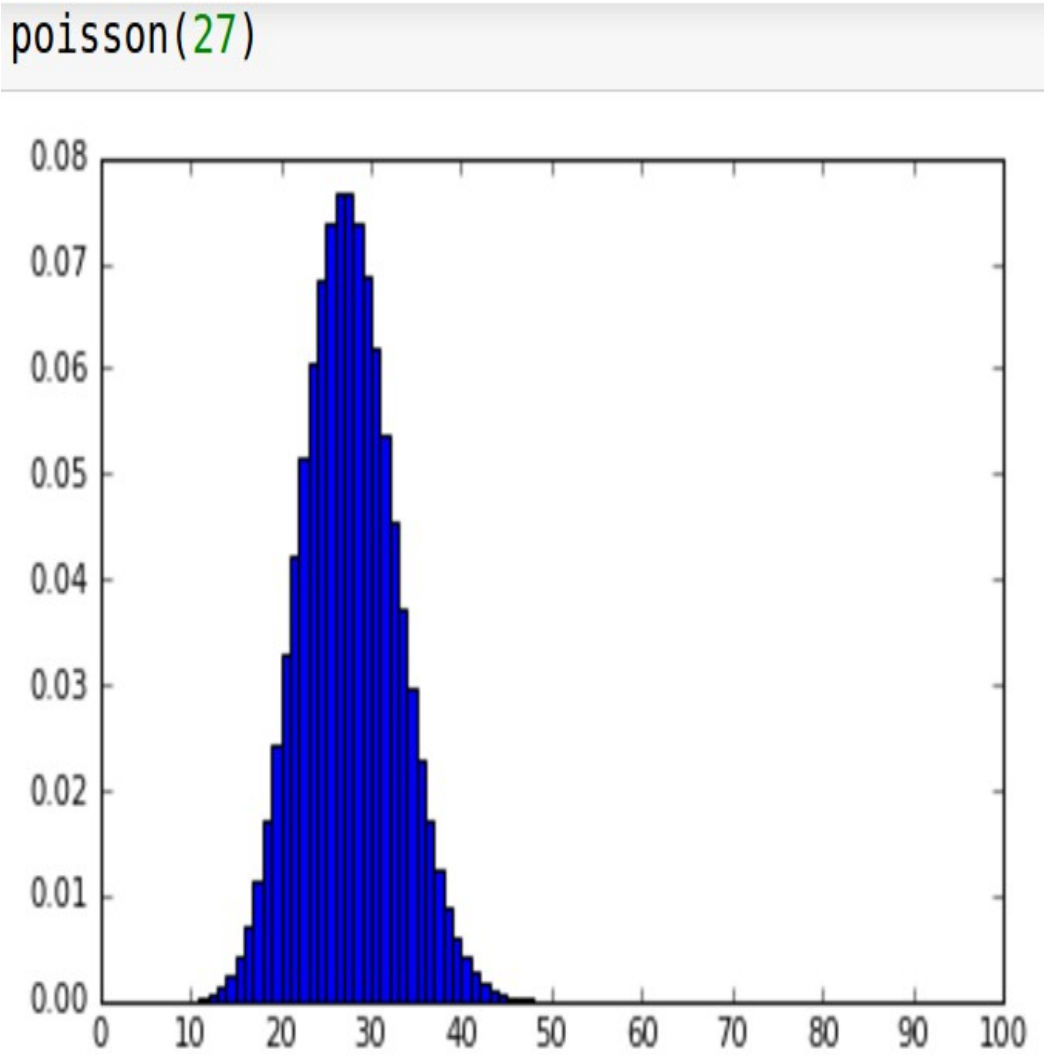
Example :

The no of hits on a website follows Poisson distribution with a mean of 27 hits per hour.

Find the probability there will be 90 or more hits in 3 hours.

Solution :
$\lambda = 27$

X -denote no of hits in three hours.
X ~ Poisson(27 * 3)
X ~ Poisson(81)
=> X ~ N(81, 9^2)



Example :

$X \sim N(81, 9^2)$

$Z = (90 - 81) / 9 = 1.00$

$P(Z >= 1.00) = 0.1587$

1) Normal approximation to Binomial

Central Limit Theorem(CLT)

Prof. Preet Kanwal

Central Limit Theorem

Let X1, X2 .. Xn be a simple random sample from a population with mean μ and standard deviation σ .

Let X_bar deonte the Sample mean

Let S_n denote sum of the sample observations

Then if n is sufficiently large then,

$$\text{X\_bar} \sim N(\mu , \sigma^2/n) \text{ approx.}$$

&

$$S_n \sim N(n\mu , n\sigma^2) \text{ approx.}$$

Problem 1

A simple random sample of 100 men is chosen from a population with mean height 70 in. & standard deviation 2.5 in.

What is the probability that the average height of the sample men is greater than 69.5 in?

Problem 1 - Solution

n = 100

Let X denote the average height of the men.

mean of sample = $\mu_x = 70$

standard deviation of sample = s = 2.5

Using CLT, we can say,

$$X\_bar \sim N(\ \mu\ ,\ \sigma^2/n\ )\ approx.$$

Since n is large, we replace μ with sample mean μ_x and population SD σ with sample SD s.

$$X\_bar \sim N(\ \mu_x\ ,\ s^2/n\ )\ approx.$$

P(X_bar > 69.5) = ?

We can standardize X_bar as z = (69.5 – 70)/(2.5/sqrt(100)) = -2

P(Z > -2) = .9772

Problem 2

The manufacture of a certain part requires two different machine operations.

- The time on machine 1 has mean 0.5 hours and standard deviation 0.4 hours.
- The time on machine 2 has mean 0.6 hours and standard deviation 0.5 hours.
- The times needed on the machines are indpendent.
- Suppose 100 parts are manufactured.

  1) What is the probability that the total time used by machine 1 is greater than 55 hours?

  2) What is the probability that the total time used by machine 2 is less than 55 hours?

  3) What is the probability that the total time used by both the machines together is greater than 115 hours.

  4) What is the probability that the total time used by machine 1 is greater than the total time used by machine 2?

Problem 2 - Solution

1) What is the probability that the total time used by machine 1 is greater than 55 hours? Let X represent the time on machine 1.

$S_x \sim N(100 * 0.5 , 100 * 0.4^2)$

$P(S_x > 55) = ?$

$P(Z > (55 – 50)/ 4) = P(Z > 1.25) = 0.1056$

Problem 2 - Solution

2) What is the probability that the total time used by machine 2 is less than 55 hours?

Let Y reprsent time on machine 2.

$S_y \sim N(100 * 0.6 , 100 * 0.5^2)$

$P(S_y < 55) = ?$

$P(Z < (55 – 60)/ 5) = P(Z < -1) = 0.1587$

Problem 2 - Solution

3) What is the probability that the total time used by both the machines together is greater than 115 hours.

Let T reprsent time on both machine 1 and 2.

$S_T \sim N((100 * 0.5 + 100 * 0.6), (100 * 0.4^2 + 100 * 0.5^2))$

$S_T \sim N(110, 41)$

$P(S_T > 115) = ?$

$P(Z > (115 - 110)/ 6.40) = P(Z > 0.78) = 0.2177$

Problem 2 - Solution

4) What is the probability that the total time used by machine 1 is greater than the total time used by machine 2?

Let D reprsent difference in time of machine 1 and 2.

$S_D \sim N((100 * 0.5 - 100 * 0.6), (100 * 0.4^2 + 100 * 0.5^2))$

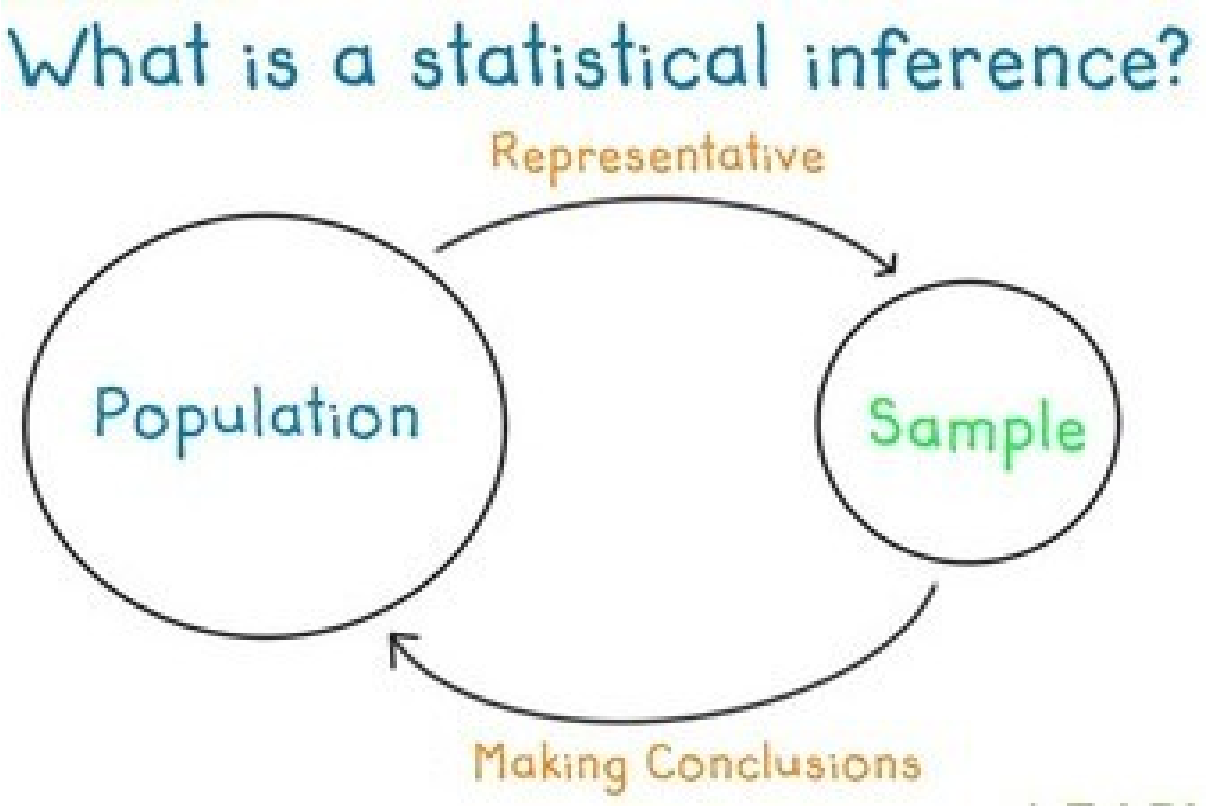
$S_D \sim N(-10, 41)$

$P(S_D > 0) = ?$

$P(Z > (0 – (-10))/ 6.40) = P(Z > 1.56) = 0.0594$

Point Estimate

Prof. Preet Kanwal



- Statistics are used to estimate the value of the parameters.
- For example:

  a) Sample Mean used to estimate Pop mean

  b) Sample proportion used to estimate Pop proportion
- These estimates are called Point estimates as Its a single value.

Example

Point estimate:

X ~ Bin(n, p) and X is observed as 7,

where n =20 then,

p_hat = 7/ 20 (**point estimate**)

p_hat = X/n (**point estimator** – no particular value of X is specified)

We address 2 Questions in this section

1) How to determine goodness of point estimator?

2) What methods can be used to construct good point estimators?

Measuring goodness of an Estimator

- An estimator must be both accurate(measured by bias) and precise(measured by uncertainity).
- MSE – Mean squared Error – quantity used to evaluate overall goodness of an estimator.

$$MSE = Bias^2 + Uncertainity^2$$

Problem 1

Let X ~ Bin(n,p) where p is unknown. Find MSE of p_hat = X / n.

Problem 2

Let X1 and X2 be independent, each with unknown mean μ and variance = 1.

1) Let μ_hat = (X1 + X2) / 2. Find bias, variance, and MSE of μ_hat.

2) Let μ_2_hat = (X1 + X2) / 4. Find bias, variance, and MSE of μ_2_hat.

3) For what values of μ does μ_2_hat have smaller MSE than μ_hat?

Method to construct good Estimator

Method of Maximum Likelihood is regarded as the **best method to point estimation.**

Likelihood function is the probability of obtaining observed value.

- If single observation is made, Likelihood fuction is just the probability of obtaining that value. (There will be no product involved)

MLE (maximum likelihood estimator) – is that value of the estimator which when substituted in place of the parameter, maximizes the Likelihood function.

MLE – when there's just one parameter to estimate

- General method:

  Step I : Write down the likelihood function.

  Step II : Take natural log of likelihood function.

  (reason: the quantity that maximizes log of a function is always the same quantity that maximizes the function itself)

  Step III : Differentiate log-likelihood fuction w.r.t parameter being estimated.

  StepIV : Set the derivative equal to 0 to get MLE.

MLE – when there's more than one parameter to estimate

- General method:

  Step I : Write down the likelihood function.

  Step II : Take natural log of likelihood function.

  (reason: the quantity that maximizes log of a function is always the same quantity that maximizes the function itself)

  Step III : Partially Differentiate log-likelihood fuction w.r.t each parameter being estimated.

  StepIV : Set the derivative in each case equal to 0 to get MLE of required parameter.

Problems

1) X ~ Bin(20, p) where X is observed as 7. Find MLE of p.

[Hence find MLE of p if X ~ (n,p)]

2) Suppose no of claims in a month from a certain portfolio of policies has a Poisson distribution with mean μ.

- The no of claims observed in one particular month is 8. Find MLE of λ.
- In one month 8 claims were incurred and in next month 10 claims were incurred. Find MLE of λ.

[Hence find MLE of λ if X ~ Poisson(λ)]

Problems

3) Find MLE of p if X ~ Geom(p).

4) Find MLE of λ of if X ~ Exp(λ).

5) Let X1, X2, Xn be a random sample from N(μ, 1) population. Find MLE of μ.

6) Let X1, X2, Xn be a random sample from N(0, σ^2) population. Find MLE of σ.

7) 6) Let X1, X2, Xn be a random sample from N(μ, σ^2) population. Find MLE of μ and σ.

CONFIDENCE INTERVALS

Prof. Preet Kanwal

Confidence Intervals

- Its an interval estimate for a population parameter.
- Based on sample data and provides a range of plausible values for a parameter.
- Confidence Interval differs from sample to sample (taken from same population).
- Associated with each Confidence Interval is a Confidence level.
- For example: we may be 95% confident that μ (population parameter) lies in the interval (-0.2, 3.1). Here, being 95% is the confidence level associated with the Confidence Interval (-0.2, 3.1).

Contents

In this unit we'll study construction of Confidence Intervals for:

a) Population Mean -

- Population mean of Large samples.
- Population mean of Small samples.
- Difference between Two Population means of Large samples.

b) Confidence Intervals with Paired Data

c) Population Proportion -

- Proportions of Large Samples.
- Difference between Two proportions of Large Samples.

Construction of Confidence Intervals for Population Mean of Large Samples:

Prof. Preet Kanwal

Construction of Confidence Intervals for Population Mean of Large Samples:

point estimate ± Margin of error

Confidence Interval for μ will be of the form:

X_bar ± Margin of error

Construction of Confidence Intervals for Population Mean of Large Samples:

A $(1 - \alpha)$ 100% Confidence Interval for μ is given by

$$X\_bar \pm z_{\alpha/2} (\sigma/\sqrt{n})$$

where, the quantity $z_{\alpha/2} (\sigma/\sqrt{n})$ is the Margin of error.

Note: $(\sigma/\sqrt{n})$ is the standard deviation of Sampling distribution of sample mean (X_bar).

Problem 1

Problem 1 : Find the value of $z_{\alpha/2}$ to use to construct a confidence interval with level:

a) 95%

b) 98%

c) 99%

d) 80%

Problem 1 : Solution

a)95% : $\bar{X} \pm 1.96\,(\sigma/\sqrt{n})$

b) 98% : $\bar{X} \pm 2.33\,(\sigma/\sqrt{n})$

c) 99% : $\bar{X} \pm 2.57\,(\sigma/\sqrt{n})$

or

$\bar{X} \pm 2.58\,(\sigma/\sqrt{n})$

d) 80% : $\bar{X} \pm 1.28\,(\sigma/\sqrt{n})$

Problem 2

Find the levels of confidence intervals that have the following values of $z_{\alpha/2}$:

a) $z_{\alpha/2} = 2.17$

b) $z_{\alpha/2} = 3.28$

Problem 2 : Solution

a) $z_{\alpha/2} = 2.17$

$$P(-z_{0.015} < Z < z_{0.015}) = 1 - 0.03$$

Hence the confidence level is 97%

b) $z_{\alpha/2} = 3.28$

$$P(-z_{0.0005} < Z < z_{0.0005}) = 1 - 0.001$$

Hence the confidence level is 99.9%

Problem 3

In a sample of 100 wires the average breaking strength is 50kN, with a satandard deviation of 2kN.

a) Find 68% confidence interval fzor the mean breaking strength of this type of wire.

b) Find 80% confidence interval for the mean breaking strength of this type of wire.

c) Find 90% confidence interval for the mean breaking strength of this type of wire.

d) Find 95% confidence interval for the mean breaking strength of this type of wire.

e) Find 99% confidence interval for the mean breaking strength of this type of wire.

Problem 3 : Solution

Confidence level	Margin of error	Width of interval
68	0.2	(49.8, 50.2)
80	0.256	(49.744, 50.256)
90	0.33	(49.67, 50.33)
95	0.392	(49.608, 50.392)
99	0.516	(49.484, 50.516)

Trade of between Confidence level we pick and the width of the interval. This means higher the confidence level, wider the interval.

Problem 3 - Continued

In a sample of 100 wires the average breaking strength is 50kN, with a satandard deviation of 2kN.

f) An engineer claims that the mean breaking strength is between 49.7kN and 50.3kN. With what level of confidence can this statement be made?

g) How many wires must be sampled so that a 95% confidence interval specifies the mean breaking strength to within 0.3 kN?

h) How many wires must be sampled so that a 99% confidence interval specifies the mean breaking strength to within 0.3 kN?

Problem 3 : Solution

f) An engineer claims that the mean breaking strength is between 49.7kN and 50.3kN. With what level of confidence can this statement be made?

confidence interval (49.7, 50.3) corresponds to 86.64% confidence level.

g) How many wires must be sampled so that a 95% confidence interval specifies the mean breaking strength to within 0.3 kN?

171 wires must be sampled so that a 95% confidence interval specifies the mean breaking strength to within 0.3 kN.

h) How many wires must be sampled so that a 99% confidence interval specifies the mean breaking strength to within 0.3 kN?

Hence 296 wires must be sampled so that a 99% confidence interval specifies the mean breaking strength to within 0.3 kN.

Problem 4

An investigator computes 95% confidence interval for a population mean on the basis of a sample of size 70. If she wishes to compute a 95% confidence interval that is half as wide, how large a sample does she need?

Problem 4 : Solution

Sample size should be 280.

Problem 5

A 95% confidence interval for a population mean is computed from a sample of size 400. Another 95% confidence interval will be computed from a sample of size 100, drawn from the same population. Choose the best answer to fill in the blank; the interval from the sample of size 400 will be approximately _____ as the interval from the sample of size 100.

a) 1/8th as wide b) 1/4th as wide c) 1/2 as wide d) same width

e) twice as wide f) four times as wide h) eight times as wide

Problem 5 - Solution

The ratio of the widths is equal to the ratio of the standard deviations of the sample mean,

$$1.96(\sigma/20) : 1.96(\sigma/10) \Rightarrow (\sigma/20) : (\sigma/10) \Rightarrow 10 : 20 \Rightarrow 1:2$$

=>the interval from the sample of size 400 will be approximately **half(1/2) as wide** as the interval from the sample of size 100. **[option c]**

Interpreting Confidence Intervals

Prof. Preet Kanwal

Assumptions made in interpreting a CI of a mean:

To interpret the confidence interval of the mean, you must assume that:

All the values were **independently and randomly sampled** from a population whose values are distributed according to a **Gaussian(Normal) distribution.**

Probability vs Confidence

It is correct to say that there is a 95% chance that the confidence interval you calculated contains the true population mean.

It is not quite correct to say that there is a 95% chance that the population mean lies within the interval.

The population mean has one value.

In contrast, the confidence interval you compute depends on the data you happened to collect.

Meaning and Interpretaion of Confidence Interval:

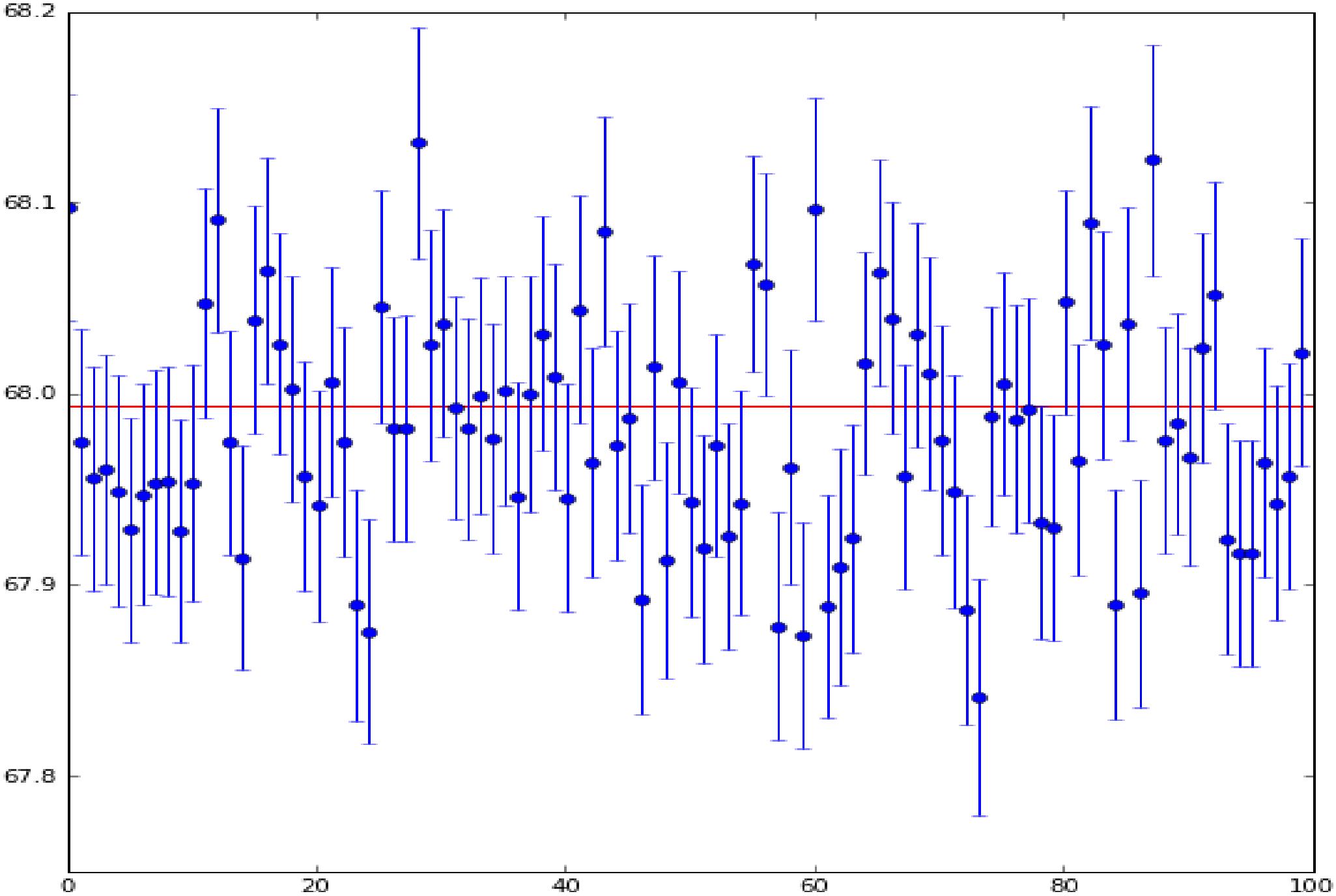
Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 90% of the time.

Python Demo : Use height-weight csv.

Take 100 samples each of sample size 1000.

Compute 95% CI for each sample and observe how many of CI's contain μ

68% Confidence Level



Misinterpretations of Confidence Intervals

a) A 95% confidence interval does not mean that for a given realised interval calculated from sample data there is a 95% probability the population parameter lies within the interval, nor that there is a 95% probability that the interval covers the population parameter.

b) A 95% confidence interval does not mean that 95% of the sample data lie within the interval.

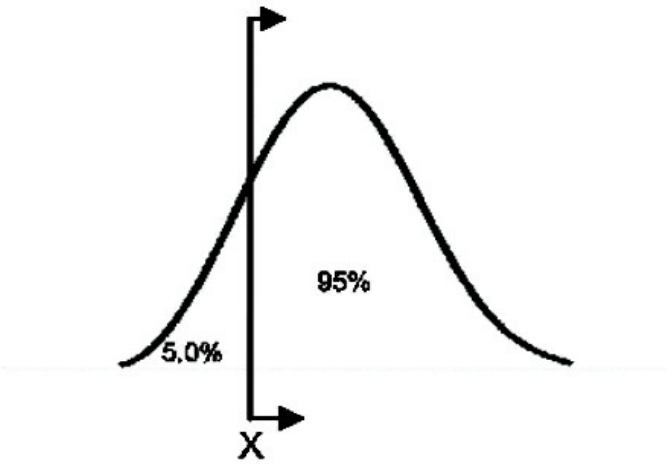
c) A confidence interval is not a range of plausible values for the sample mean, though it may be understood as an estimate of plausible values for the population parameter.

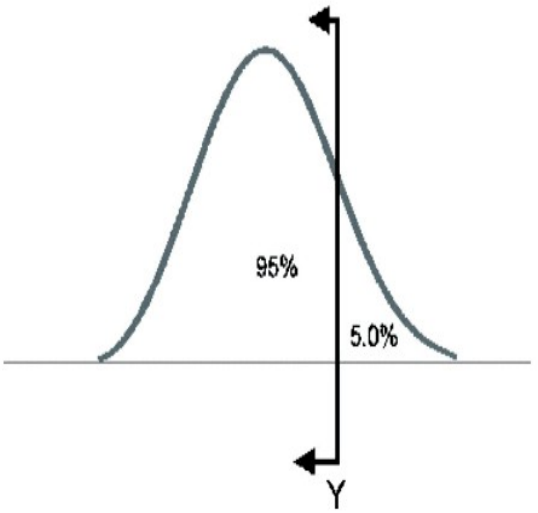
One-sided Confidence Intervals

1) An upper one-sided bound defines a point that a certain percentage of the population is less than. For example, if X is a 95% upper one-sided bound, this would indicate that 95% of the population is less than X.

2) A lower one-sided bound defines a point that a specified percentage of the population is greater than.

For example, If X is a 95% lower one-sided bound, this would indicate that 95% of the population is greater than X.





Let $X_1, \ldots, X_n$ be a *large* ($n > 30$) random sample from a population with mean μ and standard deviation σ, so that $\overline{X}$ is approximately normal. Then level $100(1-\alpha)\%$ lower confidence bound for μ is

$$\overline{X} - z_\alpha \sigma_{\overline{X}} \tag{5.2}$$

and level $100(1-\alpha)\%$ upper confidence bound for μ is

$$\overline{X} + z_\alpha \sigma_{\overline{X}} \tag{5.3}$$

where $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When the value of σ is unknown, it can be replaced with the sample standard deviation s.

Problem 1

In a sample of 80 ten-penny nails, the average weight was 1.56g and the standard deviation was 0.1g.

a) Find a 90% upper confidence bound for the mean weight.

b) Find a 80% lower confidence bound for the mean weight.

c) Someone says that the mean weight is less than 1.585g. With what level of confidence can this statement be made?

Problem 1 : Solution

a) 90% upper confidence bound for the mean weight.

= 1.5743

b) Find a 80% lower confidence bound for the mean weight.

= 1.551

c) Someone says that the mean weight is less than 1.585g. With what level of confidence can this statement be made?

Hence we can make the statement with 98.75% confidence.

Problem 2

One step in the manufacture of a certain metal clamp involves the drilling of four holes. In a sample of 150 clams, the average time needed to complete this step was 72 seconds and the standard deviation was 10 seconds.

An efficiency expert says that the mean time is greater than 70 seconds. With what level of confidence can this statement be made?

Problem 2 : Solution

mean = 72

sigma = 10

n = 150

Mean > 70

That means the lower confidence bound = 70

mean – lower_bound = 72 - 70 = 2

=> - z * (10/sqrt(150)) = 2

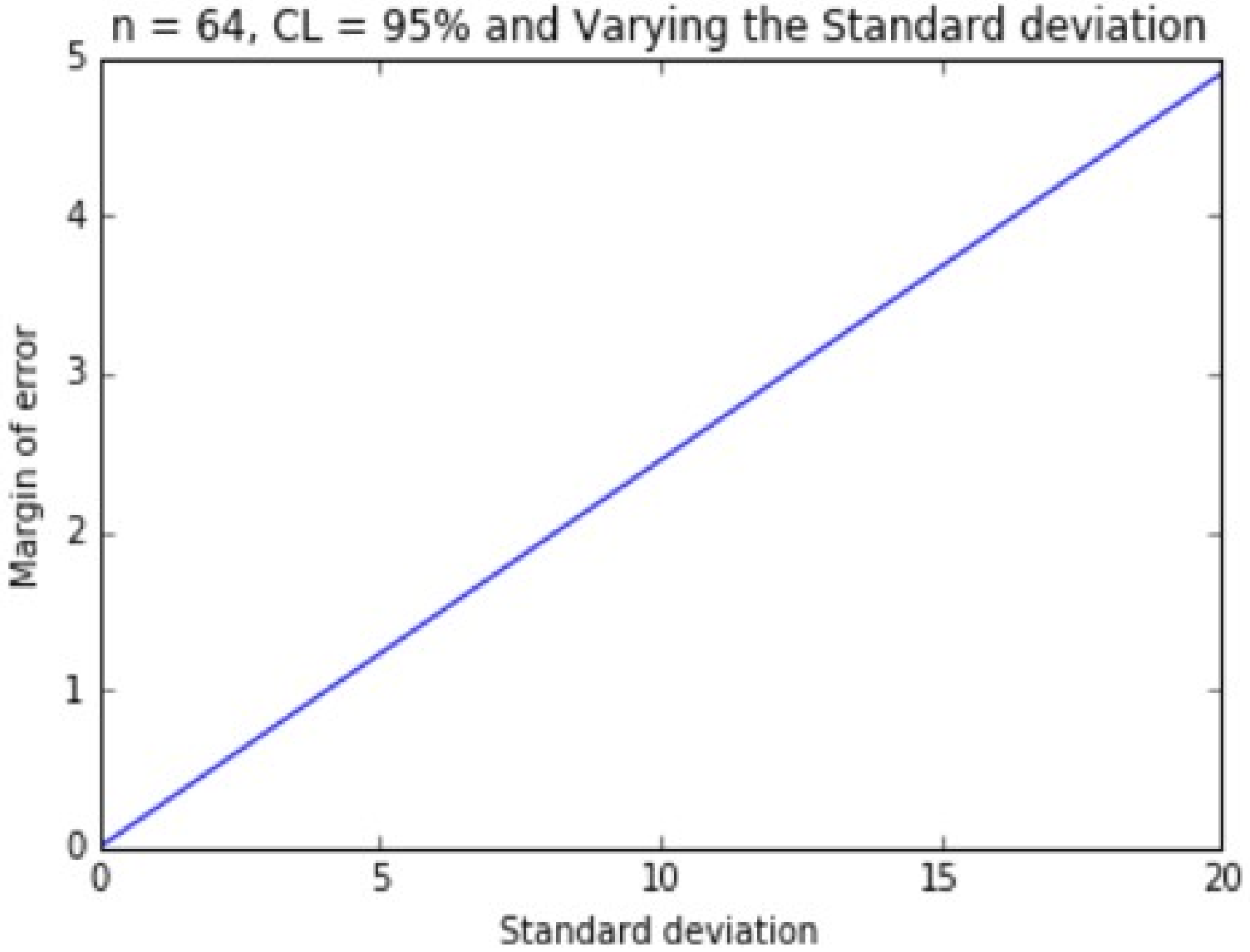
=> z = - 2.449 = - 2.45

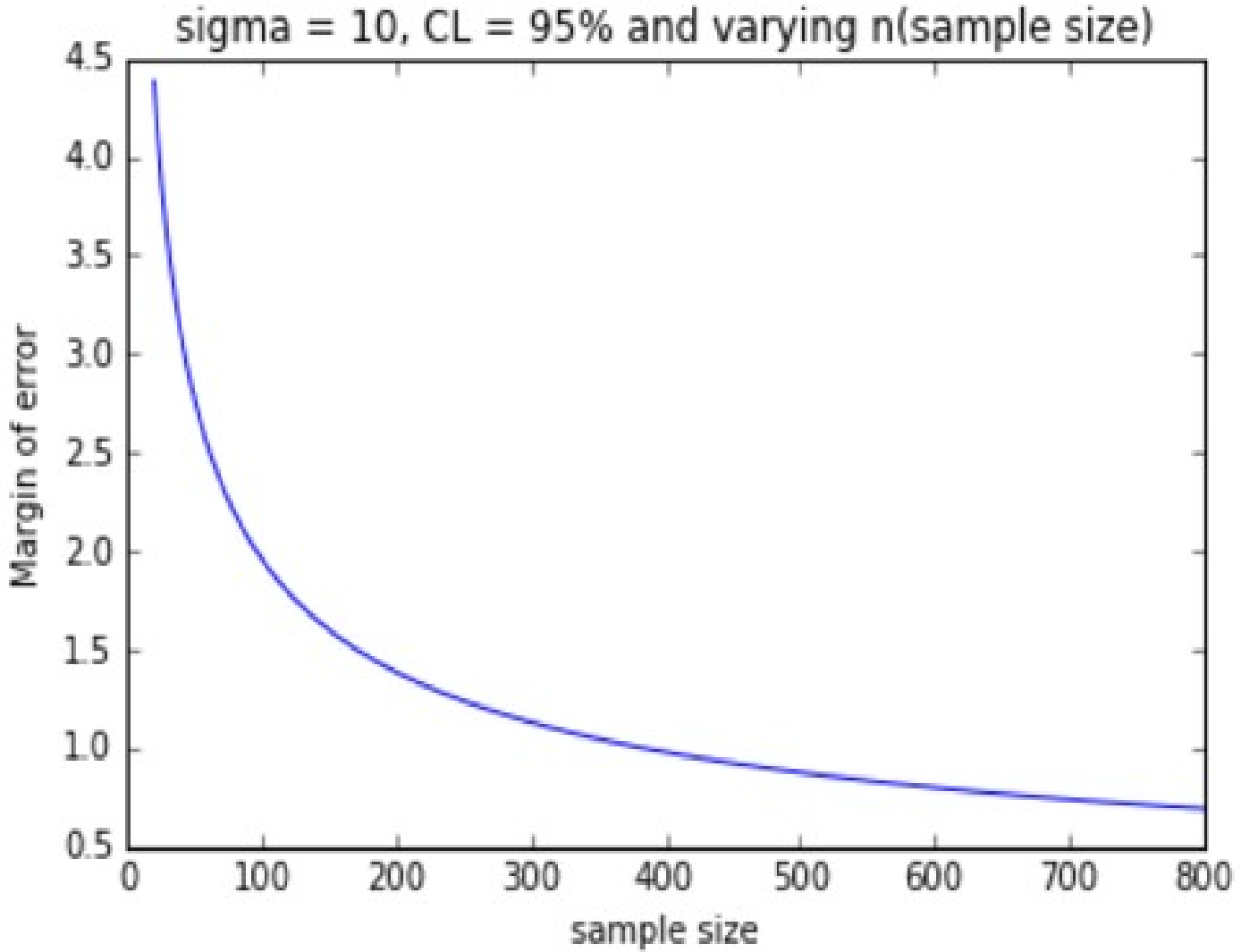
=> Area to right of -2.45 = Area to the left of 2.45 = 0.9929

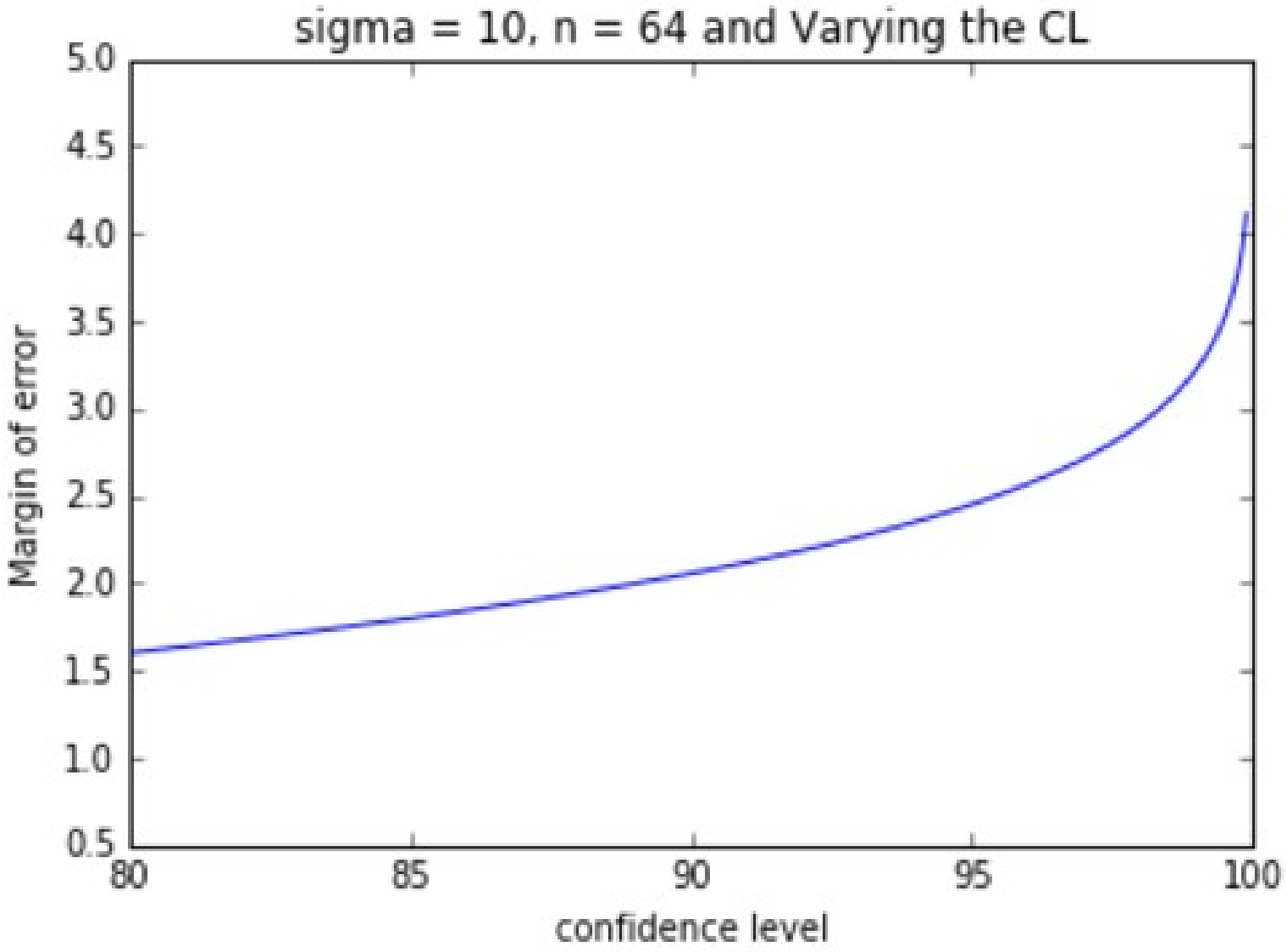
Hence we can make the statement with 99.29% confidence.

Factors that affect the Margin of Error

$\left(z_{\alpha/2} (\sigma / \sqrt{n}) \right)$


n = 64, CL = 95% and Varying the Standard deviation
Margin of error
Standard deviation
0
1
2
3
4
5
0
5
10
15
20


sigma = 10, CL = 95% and varying n(sample size)
Margin of error
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
100
200
300
400
500
600
700
800
sample size


sigma = 10, n = 64 and Varying the CL
Margin of error
confidence level
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
80
85
90
95
100

Observations

1) **Margin of error increases linearly with increase in standard deviation**(as Sigma is in the numerator) => keeping all other factors constant, doubling the SD doubles the Margin of error.

2) The **greater the Sample size**, more information we have and more precisely we'll be able to pin down the value of the parameter of interest and **smaller the margin of error.** Hence in order to cut the margin of error in half, we would need to quadruple the sample size.

3) We know, there's a trade off between the confidence level and the width of the interval. **Greater the confidence, greater the margin of error. Beyond 95% CL the curve goes up quickly**, hence in most practical situations we feel **95% CL provides a good balance between high confidence and reasonable margin of error.**

Construction of Confidence Intervals for Population Mean of Small Samples (n < 30)

Prof. Preet Kanwal

Introduction

If the sample size is small, standard deviation (s) of the sample may not be close to σ (population standard deviation). Hence X_bar(sample_mean) may not be aproximately normal.

However, if the population from which the sample is drawn is known to be approximately normal (can be confirmed using normal probability plot). It turns out that we can still use the quantity $(\bar{X} - \mu) / (s/\sqrt{n})$, but since s is not necessarily close to σ , the quantity will not have a normal distribution.

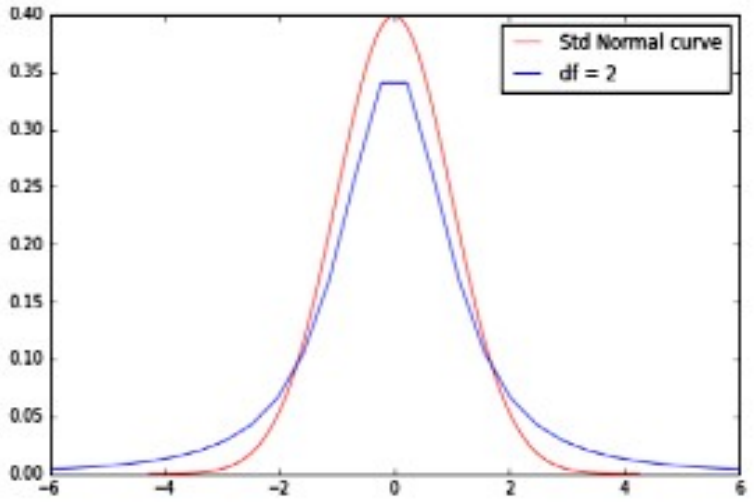
Instead it has Student's t distribution with n – 1 degrees of freedom, denoted as t_{n-1} .

THE t DISTRIBUTION

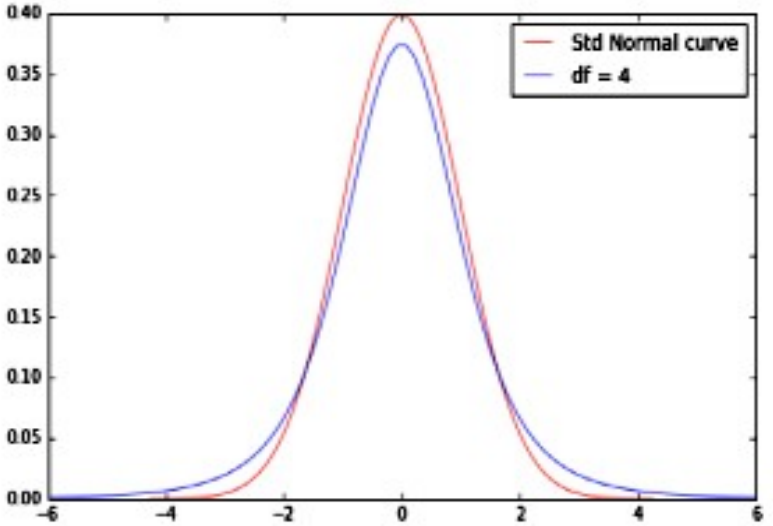
- The t distribution is a theoretical probability distribution.
- It is symmetrical, bell-shaped, and similar to the standard normal curve.
- It differs from the standard normal curve, however, in that it has an additional parameter, called **degrees of freedom**, which changes its shape.
- **df = sample size – 1**
- Setting the value of df defines a particular member of the family of t distributions.(df > 0 => Sample Size > 1)

Student's t distribution

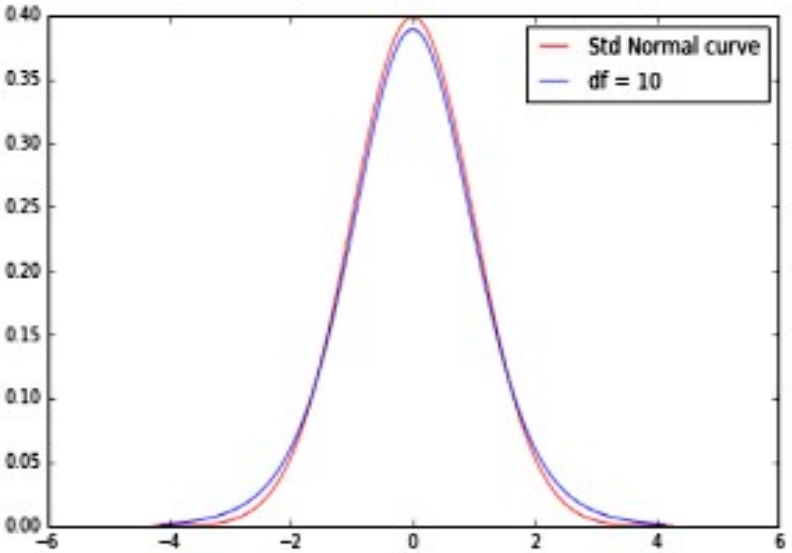
1) df = 2



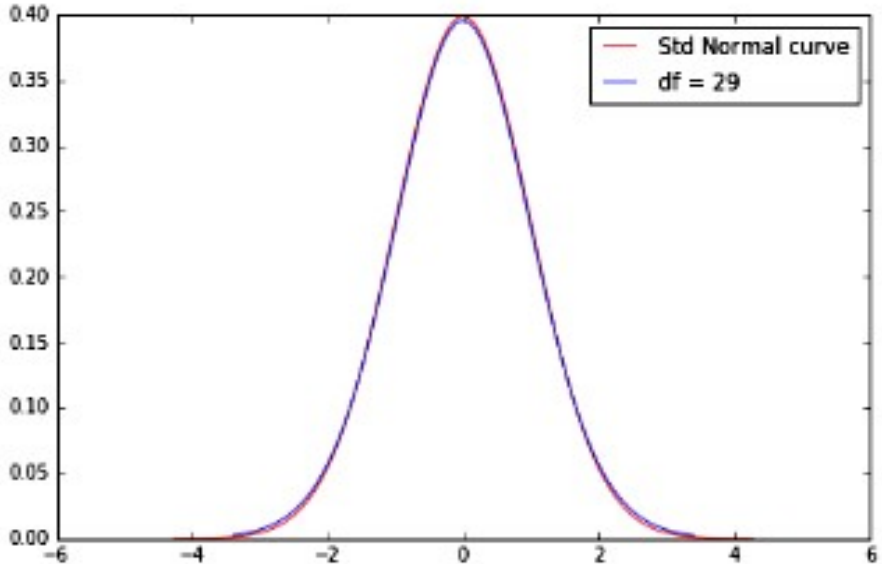
2) df =4



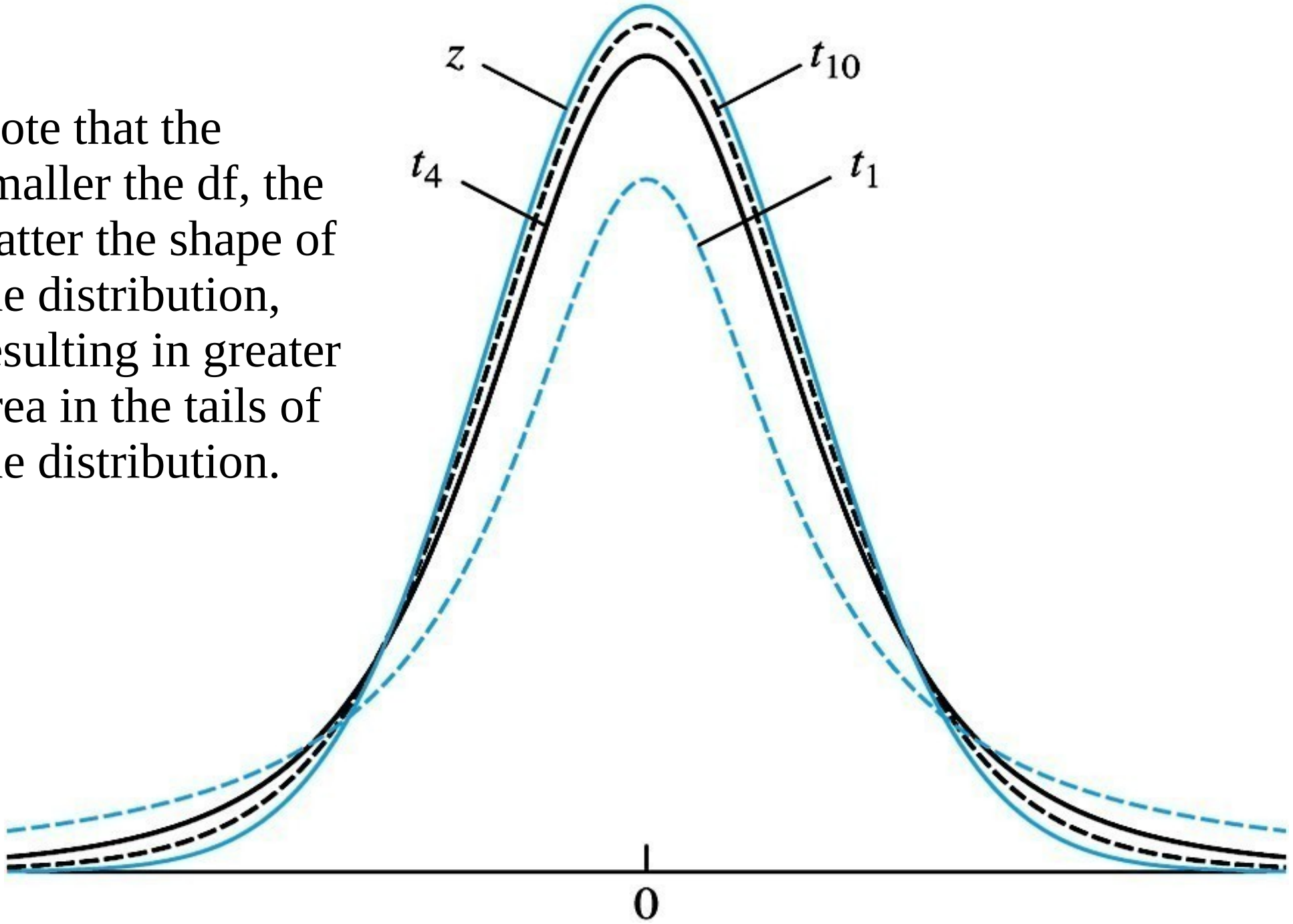
3) df = 10



4) df = 30



Note that the smaller the df, the flatter the shape of the distribution, resulting in greater area in the tails of the distribution.



RELATIONSHIP TO THE NORMAL CURVE

- As the df increase, the t distribution approaches the standard normal distribution (μ=0.0, σ=1.0).
- The standard normal curve is a special case of the t distribution when df= infinity.
- For practical purposes, the t distribution approaches the standard normal distribution relatively quickly, such that when df=30 the two are almost identical.

Using t table

We use t table to find probabilites associated with t distribution.

Row headings – denotes degree of freedom

Column headings – denotes the area to the right(probabilities)

The value in particular row and column specifies the t-score where,

P(t > t-tscore) = col_heading

Problem 1

1) A random sample of size 10 is drawn from a normal distribution with mean 4.

a) Find P(t >1.833)

b) Find P(t > 1.5)

2) Find the value of t_{12} distribution where upper-tail probability is 0.025.

Solution: Problem 1 – Part 1 (a)

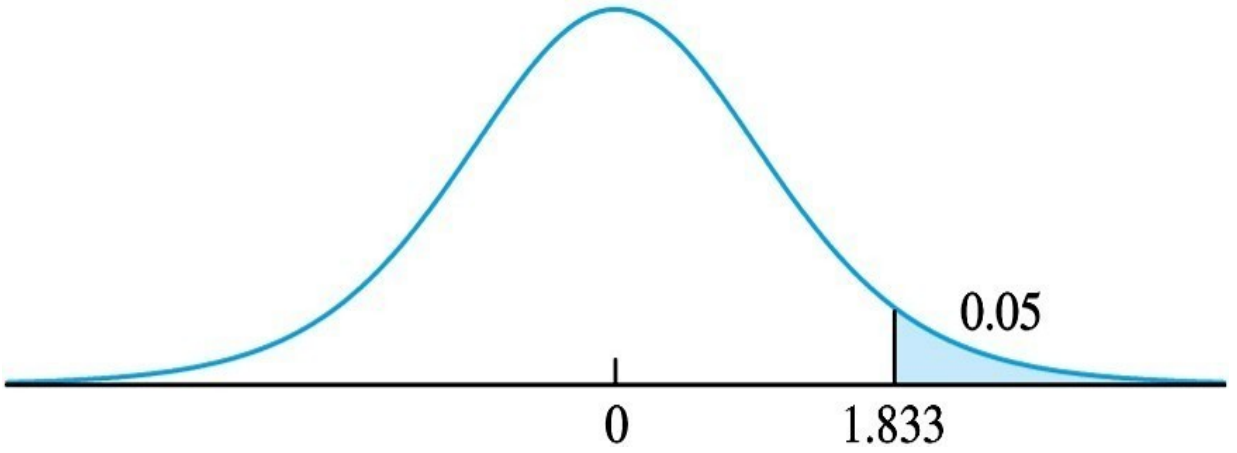
a) Find P(t >1.833)

df = 9 (row_heading)

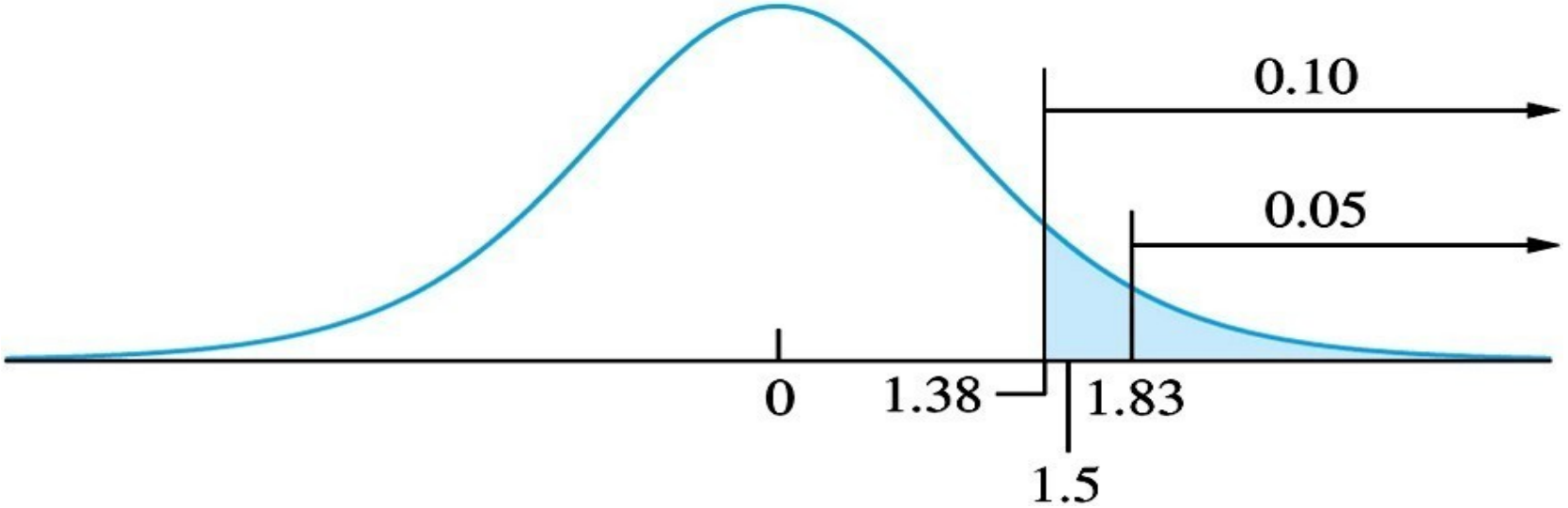
t-score = 1.833

corresponding col_heading = 0.05

P(t >1.833) = 0.05



Solution: Problem 1 – Part 1 (b)



b) Find P(t > 1.5)

df = 9 (row_heading)

t-score = 1.5 [does not correspond to any of the values in that row]

but we do have t-scores 1.383, 1.833 corresponding to upper tail probabilties 0.10 and 0.05 respectively. That is,

P(t >1.383) = 0.10 and P(t >1.833) = 0.05

Since 1.383 < 1 .5 < 1.833 => 0.05 < P(t > 1.5) < 0.10

Solution: Problem 1 – Part 2

Problem 2: Find the value of t_{12} distribution where upper-tail probability is 0.025.

Solution:

row_head = 12

col_head = 0.025

=> t-score = 2.179

Student's t Distribution is Appropriate when

- Sample size is small ($n < 30$)
- Sample comes from a population that is approximately normal.
- In many cases, we must examine the sample for normality, by constructing a box plot or normal probbaility plot.
- Should not be used for samples that contain outliers.

Constructing Confidence Interval for Small Samples using t distribution:

The quantity, $$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with n – 1 degrees of freedom.

We can generate a (1 – α) 100% Confidence Interval for μ as

$$\overline{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}.$$

One-Sided CI for Small Samples

We can generate a (1 – a) 100% Upper Confidence bound for μ as:

$$\bar{X} + t_{n-1,\alpha} * s/\sqrt{n}$$

We can generate a (1 – a) 100% Lower Confidence bound for μ as:

$$\bar{X} - t_{n-1,\alpha} * s/\sqrt{n}$$

Problem 2

Find the value of $t_{n-1,\alpha/2}$ needed to construct a two-sided confidence interval of the given level with the given sample size:

a) 90% with sample size 12

b) 95% with sample size 7

Problem 2 : Solution

a) 90% with sample size 12

df = 11

alpha = 0.10 => alpha/2 = 0.05

=> in t table : row_heading = 11, col_heading = 0.05 => $\mathbf{t_{11,\,0.05}} = 1.796$

b) 95% with sample size 7

df = 6

alpha = 0.05 => alpha/2 = 0.025

=> in t table : row_heading = 6, col_heading = 0.025 => $\mathbf{t_{6,\,0.025}} = 2.447$

Problem 3

Find the level of two-sided confidence interval that is based on the given value of $t_{n-1,\alpha/2}$ and the given sample size:

a) t = 5.841, sample size = 4

b) t = 1.746, sample size = 17

Problem 3 : Solution

a) t = 5.841, sample size = 4

df = 3

In t table row_heading = 3 and look for corresponding col_heading where row value = 5.841

=> col_heading = 0.005 = alpha/2 => alpha = 0.01 => confidence level = 99%

b) t = 1.746, sample size = 17

df = 16

In t table row_heading = 16 and look for corresponding col_heading where row value = 1.746

=> col_heading = 0.05 = alpha/2 => alpha = 0.1 => confidence level = 90%

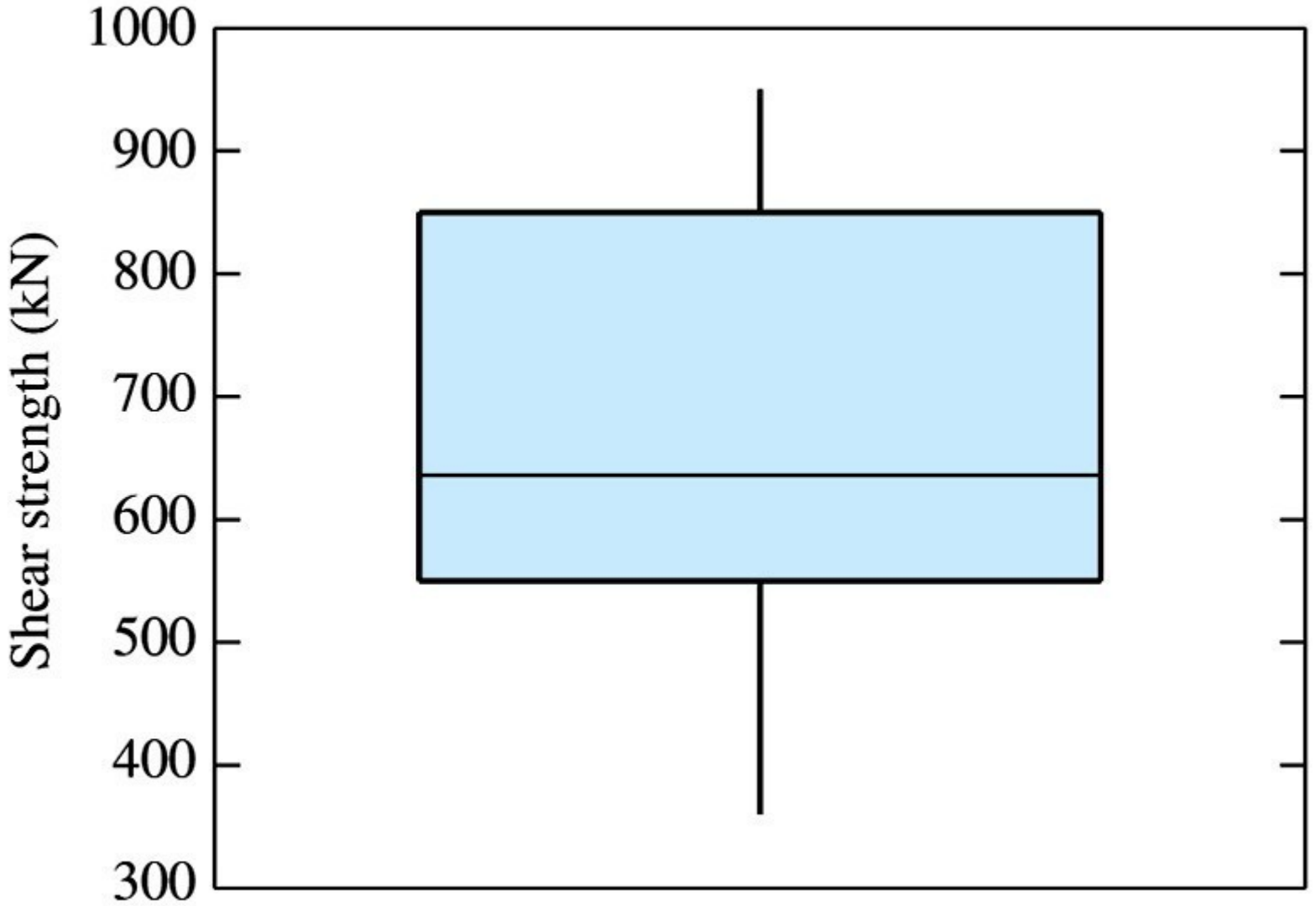
Problem 4

Following represents the measurements of the nominal shear strength (in kN) for a sample of 15 prestressed concrete beams:

580	**400**	**428**	**825**	**850**	**875**	**920**	**550**
575	**750**	**636**	**360**	**590**	**735**	**950**	

a) Is it appropriate to use the Student's t statistic to construct a 99% confidence interval for the mean shear strength?

b) If so, construct the confidence interval. If not, explain why not.



Since there are no outliers in the data set, Student's t statistic can be used to construct 99% CI.

Problem 4 – Solution

- Sample mean = X_bar = 668.27
- Sample standard deviation = s = 192.089
- $\mathbf{t_{n-1,\alpha/2} = t_{15-1,0.005} = 2.977}$
- 99% CI :

$$668.27 \pm 2.977 * 192.089/\text{sqrt}(15)$$

$$=(520.62, 815.92)$$

Problem 5

The following are summary statistics for a data set. Would it be appropriate to use the Student's t distribution to construct a confidence interval from these data?

n	mean	median	sigma	min	max	Q1	Q3
10	8.905	6.105	9.690	0.512	39.920	1.967	8.103

Problem 5 : Solution

There should be no outliers in the data in order for t distribution to be appropriate.

IQR = Q3 – Q1 = 6.136

Lower_ext = Q1 – 1.5(IQR) = Q1 – 9.204 = -7.237

Since, min > Lower_ext => there are no outliers in the negative direction.

Upper_ext = Q3 + 1.5(IQR) = Q3 + 9.204 = 17.307

Since max > Upper_ext => there are outliers in the positive direction.

Hence, t distribution is not appropriate, as the sample does not appear to come from a normal population due to the presence of an outlier.

Problem 6

The following data presents a confidence interval for a population mean, but some of the numbers are missing. Fill the missing numbers for (a) , (b) and (c).

n	mean	sigma	SE of mean	99% CI
20	2.39374	(a)	0.52640	((b) , (c))

Problem 6 : Solution

SE of mean = sigma / sqrt(n)

=> sigma = SE of mean * sqrt(n) => sigma = 0.52640 * sqrt(20) => **sigma = 2.354**

Finding 99% CI:

mean of the sample: X_bar = 2.39374

sigma (s) = 2.354

n = 20 , alpha = 0.01

=> $\mathbf{t_{19\ ,\ 0.005}} = 2.861$

=> 99% CI is:

$$X\_bar \pm t_{n-1\ ,\ \alpha/2}\ (s/\sqrt{n})$$

$$=> 2.39374 \pm 2.861\ (\ 0.52640) = 2.39374 \pm 1.506$$

Hence **99% CI is (0.88774 , 3.89974).**

Note: Use z not t if standard deviation of the population is known.

If it is known that the sample indeed was drawn from a **normal population**, also the **standard deviation of the population is known**, use z not t distribution to find out the confidence interval irrespective of the sample size.

Construction of Confidence Intervals for Difference between Two Population Means of Large Samples

Prof. Preet Kanwal

Sum/ Difference of two independent normally distributed random variables is normal

If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent random variables that are normally distributed, then their sum/difference is also normally distributed. i.e., if

If,

$$X \sim N(\mu_X, \sigma_X^2)$$

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

Then,

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

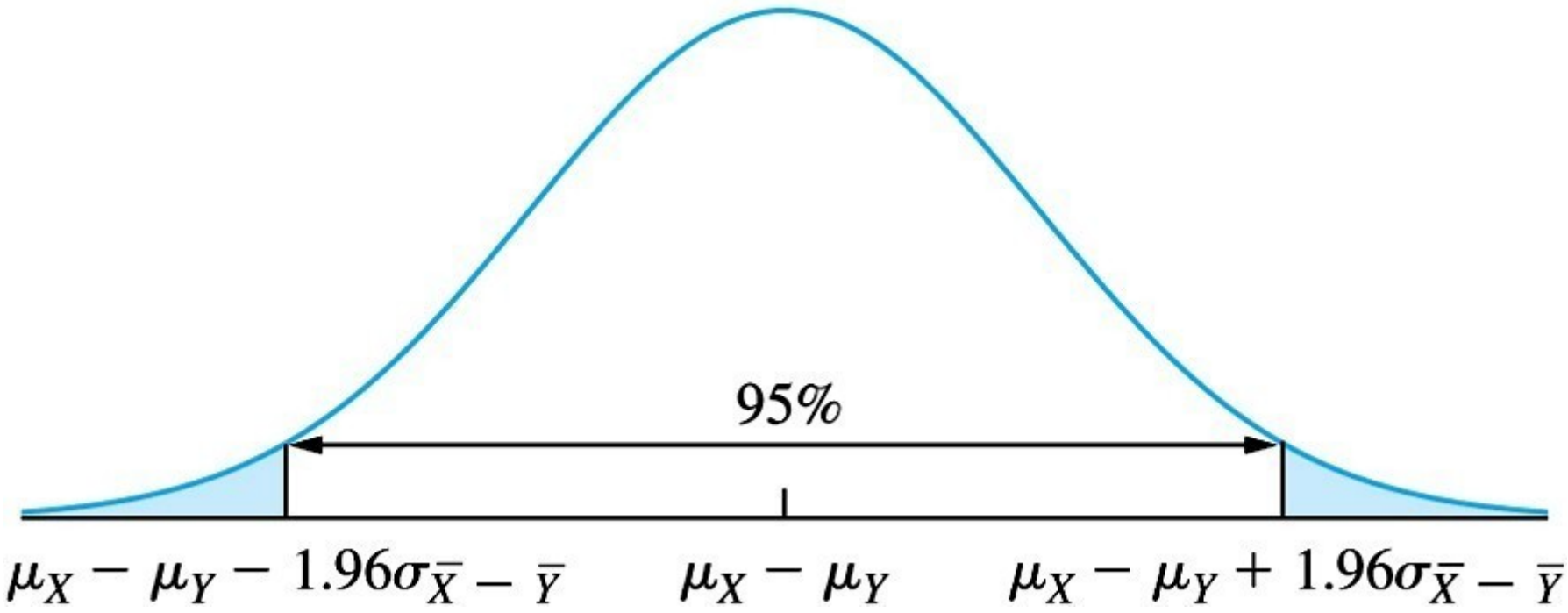
A Confidence Interval for the Difference Between Two Means

Let $X_1, \ldots, X_{n_X}$ be a *large* random sample of size n_X from a population with mean μ_X and standard deviation σ_X, and let $Y_1, \ldots, Y_{n_Y}$ be a *large* random sample of size n_Y from a population with mean μ_Y and standard deviation σ_Y. If the two samples are independent, then a level $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \qquad (5.16)$$

When the values of σ_X and σ_Y are unknown, they can be replaced with the sample standard deviations s_X and s_Y.

A Confidence Interval for the Difference Between Two Means



Problem 1

A group of 75 people enrolled in a weight loss program that involved adhering to a special diet and to a daily exercise program. After 6 months, their mean weight loss was 25 pounds, with a sample standard deviation of 9 pounds.

A second group of 43 people went on the diet but didn't exercise. After 6 months, their mean weight loss was 14 pounds, with a sample standard deviation of 7 pounds.

Find a 95% confidence interval for the mean difference between the weight losses.

Problem 1 : Solution

X_bar ~ N(25, 9/sqrt(75))

Y_bar ~ N(14, 7/sqrt(43))

since both the samples are independent,

a 95% Confidence Interval for $\mu_X - \mu_Y$ is given by

$$(X\_bar - Y\_bar) \pm z_{a/2} * sqrt((\sigma_X^2/n_1) + (\sigma_Y^2/n_2))$$

= $(25 - 14) \pm 1.96 * sqrt((9^2/75) + (7^2/43))$

= $11 \pm 1.96 * sqrt(2.2195)$

= 11 ± 2.92

= (8.08, 13.92)

Construction of Confidence Intervals for Paired Data

Prof. Preet Kanwal

Paired data

The data is described as paired when it arises from the same observational unit An example of paired data would be a before-after drug test.

The data is described as unpaired or independent when the sets of data arise from separate observational unit. For example one clinical trial might involve measuring the blood pressure from one group of patients who were given a medicine and the blood pressure from another group not given it.

Constructing Confidence Intervals with Paired data:

For large samples,

If the population of differences is approximately normal, then a

(1 – α) 100% Confidence Interval for μ_D is given by

$$\overline{D} \pm z_{\alpha/2}\sigma_{\overline{D}}.$$

In practice, $\sigma_{\overline{D}}$ is approximated with $s_D/\sqrt{n}$.

For small samples (n < 30),

If the population of differences is approximately normal, then a (1 – α) 100% Confidence Interval for μ_D is given by

$$\overline{D} \pm t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}.$$

Breathing rates, in breaths per minute were measured for a group of 10 people at rest and then during moderate exercise. The results are as follows:

N	Exercise	Rest
1	30	15
2	37	16
3	39	21
4	37	17
5	40	18
6	39	15
7	34	19
8	40	21
9	38	18
10	34	14

Find a 95% confidence interval for the increase in breathing rate due to exercise.

Solution

N	Exercise(X)	Rest (Y)	Difference (D = X – Y)
1	30	15	15
2	37	16	21
3	39	21	18
4	37	17	20
5	40	18	22
6	39	15	24
7	34	19	15
8	40	21	19
9	38	18	20
10	34	14	20

D_bar = mean of differences = 19.4

s_D = standard deviation of differences = 2.836273

n = 10 , alpha = 0.05

$t_{10-1,.025}$ = 2.262

The 95% confidence interval is 19.4 ± 2.262(2.836273/ √10), or (17.3712, 21.4288).

Construction of Confidence Intervals for Proportions of Large Samples

Prof. Preet Kanwal

Confidence Intervals for Proportions of Large samples

The method that we discussed in the last section was for a mean from any population from which a large sample has been drawn.

When the population has a Bernoulli distribution, this expression takes on a special form.

Sampling Distribution of p_hat

If X ~ Bin(n , p) where n is large. Then,

Estimate of p , p_hat has the following distribution (as follows from CLT since n is large)

p_hat ~ N(p, p(1-p)/n)

It is then also true that for 95% of all possible samples,

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

The Traditional Method – Constructing CI for proportions

Let p_hat be the proportion of successes in a *large* number of independent Bernoulli trials with success probability p.

Then the traditional level $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The method cannot be used unless the sample contains at least 10 successes and 10 failures.

Constructing CI for proportions

$X \sim \text{Bin}(n, p)$ **[n is large].** (CI : estimate ±Margin of error)

Then a $100(1-\alpha)\%$ confidence interval for p is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}.$$

Where,

$$\tilde{p} = \frac{X+2}{\tilde{n}} \quad \& \quad \tilde{n} = n+4$$

If the lower limit is less than 0, replace it with 0.

If the upper limit is greater than 1, replace it with 1.

One-sided CI for proportions

Let X be the number of successes in n independent Bernoulli trials with success probability p, so that $X \sim \text{Bin}(n, p)$.

Define $\tilde{n} = n + 4$, and $\tilde{p} = \dfrac{X+2}{\tilde{n}}$. Then a level $100(1-\alpha)\%$ lower confidence bound for p is

$$\tilde{p} - z_\alpha \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \tag{5.6}$$

and level $100(1-\alpha)\%$ upper confidence bound for p is

$$\tilde{p} + z_\alpha \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \tag{5.7}$$

If the lower bound is less than 0, replace it with 0. If the upper bound is greater than 1, replace it with 1.

Problem 1 – Part a

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

1) What proportion of the automobiles in the sample had emission levels that exceed the standard?

Solution : Problem 1 – Part a

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

1) What proportion of the automobiles in the sample had emission levels that exceed the standard?

X = 28 , n =70 => p_hat = 28/70 = 0.4

Problem 1 – Part b

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

2) Find 95% CI for the proportion of automobiles in the state whose emission levels exceed the standard.

Solution : Problem 1 – Part b

2) Find 95% CI for the proportion of automobiles in the state whose emission levels exceed the standard.

X_tilde = X + 2 = 28 + 2 = 30

n_tilde = n + 4 = 70 + 4 = 74

p_tilde = X_tilde / n_tilde = 30 / 74 = 0.405

95% CI for p:

= 0.405 ± 1.96 * sqrt (0.405 (1 – 0.405) / 74)

= 0.405 ± 0.1118

= (0.2932 , 0.5168)

Problem 1 – Part c

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

3) Find 98% CI for the proportion of automobiles in the state whose emission levels exceed the standard.

Solution : Problem 1 – Part c

3) Find 98% CI for the proportion of automobiles in the state whose emission levels exceed the standard.

X_tilde = X + 2 = 28 + 2 = 30

n_tilde = n + 4 = 70 + 4 = 74

p_tilde = X_tilde / n_tilde = 30 / 74 = 0.405

98% CI for p:

= 0.405 ± 2.33 * sqrt (0.405 (1 – 0.405) / 74)

= 0.405 ± 0.1330

= (0.272 , 0.538)

Problem 1 – Part d

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

4) How many automobiles must be sampled to specify the proportion that exceed the standard to within 0.10 with 95% confidence?

Solution : Problem 1 – Part d

4) How many automobiles must be sampled to specify the proportion that exceed the standard to within 0.10 with 95% confidence?

95% CI is given as:

$0.405 \pm 1.96 * \text{sqrt}(0.405 (1 – 0.405) / n + 4)$

$\Rightarrow 1.96 * \text{sqrt}(0.405 (1 – 0.405) / n + 4) = 0.10$

$\Rightarrow n = 88.58 \Rightarrow n = 89$

Problem 1 – Part e

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

5) How many automobiles must be sampled to specify the proportion that exceed the standard to within 0.10 with 98% confidence?

Solution : Problem 1 – Part e

5) How many automobiles must be sampled to specify the proportion that exceed the standard to within 0.10 with 98% confidence?

98% CI for p:

$0.405 \pm 2.33 * \text{sqrt} (0.405 (1 – 0.405) / n + 4)$

$\Rightarrow 2.33 * \text{sqrt} (0.405 (1 – 0.405) / n + 4) = 0.10$

$\Rightarrow n = 126.82 \Rightarrow n = 127$

Problem 1 – Part f

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

6) Find 95% lower confidence bound for the proportion of automobiles whose emissions exceed the standard.

Solution : Problem 1 – Part f

6) Find 95% lower confidence bound for the proportion of automobiles whose emissions exceed the standard.

95% lower confidence bound for p:

$0.405 – 1.645 * \text{sqrt} (0.405 (1 – 0.405) / 74)$

$= 0.3111$

Problem 1 – Part g

In a sample of 70 automobiles registered in a certain state, 28 of them were found to have emission levels that exceed a state standard.

7) Someone claims that less than half of the automobiles in the state exceed the standard. With what level of confidence can this statement be made?

Solution : Problem 1 – Part g

7) Someone claims that less than half of the automobiles in the state exceed the standard. With what level of confidence can this statement be made?

=> The upper confidence bound = 0.5

=> 0.405 + z * sqrt (0.405 (1 – 0.405) / 74) = 0.5

=> z * sqrt (0.405 (1 – 0.405) / 74) = 0.095

=> z = 1.66

=> P(Z < 1.66) = 0.9515

Hence the level is 95.15%.

Problem 2

Leakage from underground fuel tanks has been a source of water pollution. In a random sample of 87 gasoline stations, 13 were found to have at least one leaking underground tank.

a. Find a 95% confidence interval for the proportion of gasoline stations with at least one leaking underground tank.

b. How many stations must be sampled so that a 95% confidence interval specifies the proportion to within ±0.03?

Problem 2(a) : Solution

X = 13, n = 87,

- $\tilde{p} = (13 + 2)/(87 + 4) = 0.16484$,
- $z_{.025} = 1.96$.
- The confidence interval is

  $0.16484 \pm$

  $1.96 * \text{sqrt}(0.16484(1 - 0.16484)/(87 + 4))$
- (0.0886, 0.241).

Problem 2(b) : Solution

$1.96 * \text{sqrt}(0.16484(1 - 0.16484)/(n + 4)) = 0.03$

$\Rightarrow n + 4 = 587.512791$

$\Rightarrow n = 583.512791$

$\Rightarrow n = 584$

Large Samples Confidence Intervals for the Difference Between Two Proportions

Prof. Preet Kanwal

Large Samples Confidence Intervals for the Difference Between Two Proportions

Let X be the number of successes in n_X independent Bernoulli trials with success probability p_X, and let Y be the number of successes in n_Y independent Bernoulli trials with success probability p_Y, so that $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$. Define $\tilde{n}_X = n_X + 2$, $\tilde{n}_Y = n_Y + 2$, $\tilde{p}_X = (X+1)/\tilde{n}_X$, and $\tilde{p}_Y = (Y+1)/\tilde{n}_Y$.

Then a level $100(1-\alpha)\%$ confidence interval for the difference $p_X - p_Y$ is

$$\tilde{p}_X - \tilde{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{\tilde{n}_Y}} \tag{5.18}$$

If the lower limit of the confidence interval is less than -1, replace it with -1. If the upper limit of the confidence interval is greater than 1, replace it with 1.

Traditional Method

The Traditional Method for Computing Confidence Intervals for the Difference Between Proportions (widely used but not recommended)

Let $\hat{p}_X$ be the proportion of successes in a *large* number n_X of independent Bernoulli trials with success probability p_X, and let $\hat{p}_Y$ be the proportion of successes in a *large* number n_Y of independent Bernoulli trials with success probability p_Y. Then the traditional level $100(1-\alpha)\%$ confidence interval for $p_X - p_Y$ is

$$\hat{p}_X - \hat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \tag{5.19}$$

This method cannot be used unless both samples contain at least 10 successes and 10 failures.

Problem 1

In a test of the effect of dampness on electric connections:

- 100 electric connections were tested under damp conditions.
- 150 electric connections were tested under dry conditions.
- Twenty of the damp connections failed and only 10 of the dry ones failed.

Find a 90% CI for the difference between the proportions of connections that fail when damp as opposed to dry?

Problem 1 : Solution

$X = 20, n_X = 100,$

$\Rightarrow \tilde{p}_X = (20 + 1)/(100 + 2) = 0.205882$

$Y = 10, n_Y = 150,$

$\Rightarrow \tilde{p}_Y = (10 + 1)/(150 + 2) = 0.072368$

$z_{.05} = 1.645.$

Problem 1 : Solution

- The confidence interval is

  $0.205882 - 0.072368 \pm$

$$1.645\sqrt{\frac{0.205882(1-0.205882)}{100+2}+\frac{0.072368(1-0.072368)}{150+2}}$$

  = (0.0591, 0.208).

Problem 2

Angioplasty is a medical procedure in which an obstructed blood vessel is widened. In some cases, a wire mesh tube, called a stent, is placed in the vessel to help it remain open. The article “Long-term Outcomes of Patients Receiving Drug-eluting Stents” (A. Philpott, D. Southern, et al., Canadian Medical Association Journal, 2009:167–174) presents the results of **a study comparing the effectiveness of a bare metal stent with one that that has been coated with a drug designed to prevent reblocking of the vessel.**

Problem 2

- A total of 5320 patients received bare metal stents, and of these, 841 needed treatment for reblocking within a year.

- A total of 1120 received drug coated stents, and 134 of them required treatment within a year.

a) Find a 98% confidence interval for the differences between the proportions for bare metal stents and drug coated stents.

Problem 2(a) : Solution

- $X = 841$, $n_x = 5320$,

  $=> \tilde{p}X = (841 + 1)/(5320 + 2) = 0.158211$

- $Y = 134$, $n\,Y = 1120$,

  $=> \tilde{p}Y = (134 + 1)/(1120 + 2) = 0.120321$

- $z\ .01 = 2.33$.

Problem 2(a) : Solution

- The confidence interval is

$$0.158211 - 0.120321 \pm 2.33\sqrt{\frac{0.158211(1-0.158211)}{5320+2}+\frac{0.120321(1-0.120321)}{1120+2}}$$

$$= (0.0124, 0.0633).$$

Problem 2

Suppose that additional patients are to be treated in order to increase the precision of the 98% confidence interval. Three sampling plans are being considered.

- In the first plan, 1000 additional patients will be treated with bare metal stents.
- In the second plan, 500 additional patients will be treated with drug coated stents.
- In the third plan, 500 additional patients will be treated with bare metal stents and 250 additional patients will be treated with drug coated stents.

b) Which plan is most likely to provide the greatest increase in the precision of the 98% confidence interval? Explain.

Problem 2(b) : Solution

The standard deviation of the difference between the proportions is

$$\sqrt{p_X(1-p_X)/(n_X+2)+p_Y(1-p_Y)/(n_Y+2)}.$$

Estimate:

$$p_X \approx \tilde{p}_X = 0.120321 \text{ and } p_Y \approx \tilde{p}_Y = 0.158211$$

Problem 2(b) : Solution

If 1000 additional patients are treated with bare metal stents(earlier it was 5320), the standard deviation of the difference between the proportions is then :

$$\sqrt{0.120321(1-0.120321)/(1120+2)+0.158211(1-0.158211)/(6320+2)}$$

= 0.01074.

=> The width of the 98% confidence interval will be ± 2.33(0.01074) = 0.0250.

Problem 2(b) : Solution

If 500 additional patients are treated with drug-coated stents(1120 earlier), the standard deviation of the difference between the proportions is then:

$$\sqrt{0.120321(1-0.120321)/(1620+2)+0.158211(1-0.158211)/(5320+2)}$$

= 0.009502.

=> The width of the 98% confidence interval will be ±2.33(0.009502) = 0.0221.

Problem 2(b) : Solution

If 500 additional patients are treated with bare metal stents(5320 earlier) and 250 additional patients are treated with drug-coated stents(1120 earlier), the standard deviation of the difference between the proportions is then:

$$\sqrt{0.120321(1-0.120321)/(1370+2)+0.158211(1-0.158211)/(5820+2)}$$

= 0.01000.

=> The width of the 98% confidence interval will be ±2.33(0.01000) = 0.0233.

Problem 2(b) : Solution

Therefore the confidence interval would be most precise if 500 new patients are treated with drug-coated stents.

Practice Questions

Problem 1

A group of five individuals with high blood pressure were given a new drug that was designed to lower blood pressure. Systolic blood pressure was measured before and after treatment for each individual, with the following results:

Subject	Before	After
1	170	145
2	164	132
3	168	129
4	158	135
5	183	145

Find a 90% confidence for the mean reduction in systolic blood pressure.

Problem 1 : Solution

The differences are: 25, 32, 39, 23, 38.

- D_bar = 31.4
- Var = 53.3 , s_D = 7.300685
- n = 5
- $t_{5-1,.05}$ = 2.132

- The confidence interval is:

  31.4 ± 2.132(7.300685/sqrt(5)) = 31.4 ± 6.9609

  =(24.4391, 38.3609).

Problem 2

- A group of 50 computer science students were taught introductory computer programming class with an innovative teaching method that used a graphical interface and drag-and-drop methods of creating computer programs. At the end of the class, 43 of these students said that they felt confident in their ability to write computer programs.
- Another group of 40 students were taught the same material using a standard method. At the end of class, 25 of these students said they felt confident.
- Assume that each class contained a simple random sample of students. Find a 99% confidence interval for the difference between the proportions of students who felt confident.

Problem 2 : Solution

$$X = 43,\ n_X = 50,\ \tilde{p}_X = (43+1)/(50+2) = 0.846154,$$
$$Y = 25,\ n_Y = 40,\ \tilde{p}_Y = (25+1)/(40+2) = 0.619048,\ z_{.005} = 2.58.$$

The confidence Interval is:

0.846154 – 0.619048

$$\pm 2.58\sqrt{\frac{0.846154(1-0.846154)}{50+2} + \frac{0.619048(1-0.619048)}{40+2}}$$

= 0.227 ± 0.2325

= (-0.0055, 0.4595)

Problem 3

A pollster plans to survey a random sample of voters in a certain city to ask whether they support an increase in property taxes to fund the construction of a new elementary school.

How many voters should be sampled to be sure that a 95% confidence interval for the proportion who favor the proposal specifies that proportion to within ±0.04?

Problem 3 : Solution

Let n be the required sample size.

- Then n satisfies the equation

  $0.04 = 1.96 * \text{sqrt}(\tilde{p}(1 - \tilde{p})/(n + 4)$.

- Since there is no preliminary value of $\tilde{p}$ we replace $\tilde{p}$ with 0.5.

- Solving for n yields n = 596.25 => n = 597.

Problem 4

In a study of the lifetimes of electronic components, a random sample of 400 components are tested until they fail to function. The sample mean lifetime was 370 hours and the standard deviation was 650 hours. True or false:

a. An approximate 95% confidence interval for the mean lifetime of this type of component is from 306.3 to 433.7 hours.

b. About 95% of the sample components had lifetimes between 306.3 and 433.7 hours.

c. The z table can’t be used to construct confidence intervals here, because the lifetimes of the components don’t follow the normal curve.

Problem 4 : Solution

(a) True. A 95% confidence interval is found by adding and subtracting 1.96(650/ 400) from the mean of 370.

(b) False. The confidence interval specifies the location of the population mean. It does not specify the proportion of the sample items that fall into any given interval.

(c) False. So as long as the sample mean is approximately normal, which will be the case when the sample is large(CLT), the method can be used.

Thank you!