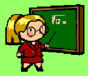
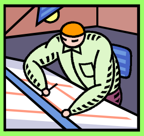
MA 3710

Why should engineers study statistics?

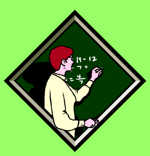

A random sample of 5 bolts yields the following measurements in inches:
{ 2.003, 1.999, 2.006, 1.996, 2.001 }

Population – the set of all people or things being studied (all possible observations)

Sample – the subset of the population from which information is actually collected

Statistics – collecting, analyzing, and interpreting data

All engineers will collect, analyze, and interpret data throughout their career.


The data set
{ 2.003, 1.999, 2.006, 1.996, 2.001 }
has variability.

An engineer would like to identify the source(s) of the variation and reduce the variation.



example

A factory is manufacturing bolts that should be exactly 2 inches long.

Measuring the length of all the bolts (a census) is too time consuming, expensive, or impossible.

Additional Information:

The bolts are being produced by two different machines:

Machine A

Machine B

Additional Information:

Each machine has three settings:

slow
500
bolts/minute

medium
1000
bolts/minute

fast
2000
bolts/minute

Machine is a <u>categorial</u> factor because its levels: A , B represent distinct settings.

Speed is a <u>continuous</u> factor because its levels: slow (500), medium (1000), fast (2000) were chosen from a continuum.

Options:

<u>Retrospective Study</u>: use data that has been collected in the past

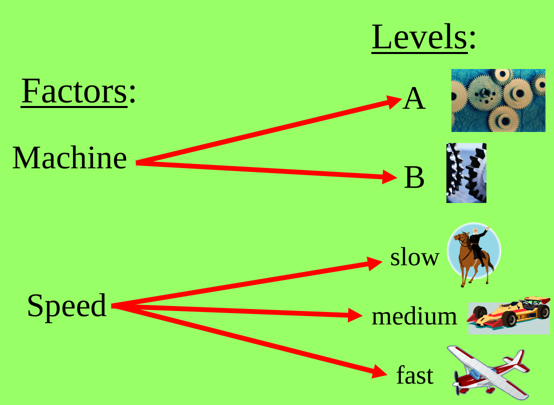
<u>Observational Study</u>: observe the process in the present and collect data

<u>Designed Experiment</u>: exhibit maximum control over when, where, and how the data is collected

There are 6 different <u>treatment combinations</u>:

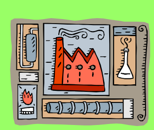
treatment	sample size	run order
Machine A - slow	10	2
Machine A - medium	10	5
Machine A - fast	10	6
Machine B - slow	10	1
Machine B - medium	10	4
Machine B - fast	10	3

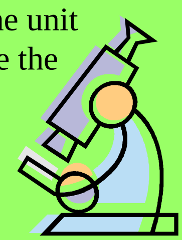
replication (sample size) randomization (run order)



<u>Experimental Unit</u>: the smallest unit to which we apply a treatment combination

<u>Observational Unit</u>: the unit upon which we make the measurement





Textbook practice problem 1.15 on page 39

A food company is considering a new recipe for their cake mix. They have identified three factors that affect texture of the cake: **amount of flour** (*low and high*), **amount of egg powder** (*low and high*), and **amount of oil added when mixing** (*low and high*). They are also interested in varying the **temperature of the oven when baking the cake** (*375 and 400*). Changing the temperature for each individual cake would be a very time-consuming process. The experiment is conducted by choosing a temperature, then mixing and baking all eight cakes together. Then the temperature is changed and again all eight cakes are baked together.

A food company is considering a new recipe for their cake mix. They have identified three factors that affect texture of the cake: **amount of flour** (*low and high*), **amount of egg powder** (*low and high*), and **amount of oil added when mixing** (*low and high*). They are also interested in varying the **temperature of the oven when baking the cake** (*375 and 400*). Changing the temperature for each individual cake would be a very time-consuming process. The experiment is conducted by choosing a temperature, then mixing and baking all eight cakes together. Then the temperature is changed and again all eight cakes are baked together.

Textbook practice problem 1.15 on page 39
c) List the levels for the factors

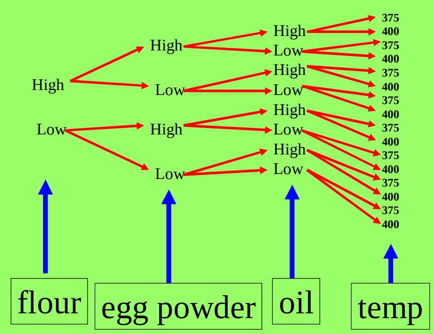
low and high (for the first three)
375 and 400 (for temperature)

A food company is considering a new recipe for their cake mix. They have identified three factors that affect texture of the cake: **amount of flour** (*low and high*), **amount of egg powder** (*low and high*), and **amount of oil added when mixing** (*low and high*). They are also interested in varying the **temperature of the oven when baking the cake** (*375 and 400*). Changing the temperature for each individual cake would be a very time-consuming process. The experiment is conducted by choosing a temperature, then mixing and baking all eight cakes together. Then the temperature is changed and again all eight cakes are baked together.

Textbook practice problem 1.15 on page 39
a) What is the response of interest?

Texture of the cake

Textbook practice problem 1.15 on page 39
d) List the possible treatment combinations



HHH375
HHH400
HHL375
HHL400
HLH375
HLH400
HLL375
HLL400
LHH375
LHH400
LHL375
LHL400
LLH375
LLH400
LLL375
LLL400

A food company is considering a new recipe for their cake mix. They have identified three factors that affect texture of the cake: **amount of flour** (*low and high*), **amount of egg powder** (*low and high*), and **amount of oil added when mixing** (*low and high*). They are also interested in varying the **temperature of the oven when baking the cake** (*375 and 400*). Changing the temperature for each individual cake would be a very time-consuming process. The experiment is conducted by choosing a temperature, then mixing and baking all eight cakes together. Then the temperature is changed and again all eight cakes are baked together.

Textbook practice problem 1.15 on page 39
b) List the factors and whether they are categorical or continuous

amount of flour, amount of egg powder, amount of oil added, and temperature of the oven (all continuous)

A food company is considering a new recipe for their cake mix. They have identified three factors that affect texture of the cake: **amount of flour** (*low and high*), **amount of egg powder** (*low and high*), and **amount of oil added when mixing** (*low and high*). They are also interested in varying the **temperature of the oven when baking the cake** (*375 and 400*). Changing the temperature for each individual cake would be a very time-consuming process. The experiment is conducted by choosing a temperature, then mixing and baking all eight cakes together. Then the temperature is changed and again all eight cakes are baked together.

Textbook practice problem 1.15 on page 39
e) Do all factors have the same experimental unit?

No, since temperature is applied to the oven, not the actual cake.

Textbook practice problem 1.15 on page 39
f) Identify the experimental unit and observational unit for temperature.
g) Identify the experimental unit and observational unit for the other factors.

f) the oven
g) a cake

This week: Read Chapter 1

MA 3710

Chapter 2 – all sections


Karlskirche
Vienna, Austria

<u>Statistics</u>

Collecting, analyzing, and interpreting data

<u>Descriptive Statistics</u>

Organizing and summarizing data

<u>example</u>

investigating lengths of all bolts being manufactured in a factory

Measure 100 bolts selected at random

One way to describe the data set:

n = 100
bolt #1 = 2.003
bolt #2 = 1.999
………
bolt #100 = 1.996

A more convenient and meaningful way to describe the data set:

average length = sample mean length =

$$\boxed{\bar{Y} = 2.001}$$

graphs and plots: histogram, dotplot, stem-and-leaf plot, boxplot

<u>Sample A</u>	<u>Sample B</u>
n = 3	n = 3
$\boxed{\bar{Y}_A = 5}$	$\boxed{\bar{Y}_B = 5}$
{ 5, 5, 5 }	{ 1, 1, 13 }
Standard Deviation = 0	Standard Deviation > 0

numerical values: mean, median, variance, standard deviation, percentiles

Inferential statistics (CH4)

Drawing conclusions from data and measuring the reliability of the conclusions

If we conducted a census we would know the population mean length = μ

Does $\bar{Y} = \mu$?

Usually not, but hopefully they are close.

$\bar{Y}$ is an estimator of μ

"Does μ = 2.000 inches ?" Yes / No

Hypothesis test

$\bar{Y} = 2.001 \rightarrow$ Does this prove that $\mu \neq 2$?

No!

Statistics rarely proves anything

$$2.001 \pm 0.003 = (1.998, 2.004)$$

estimate margin of error

Confidence interval

we are 95% confident that:

$$1.998 < \mu < 2.004$$

Histogram

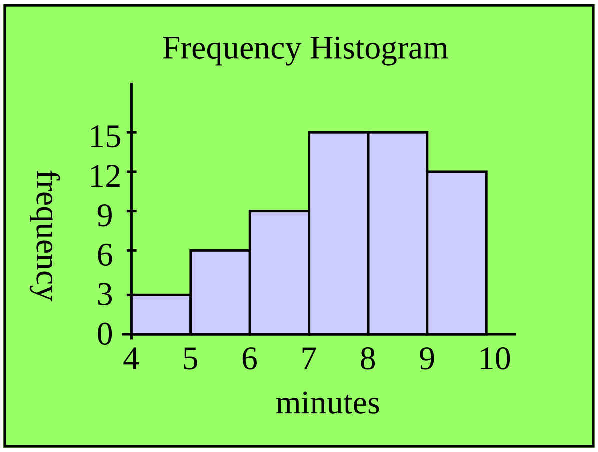
example

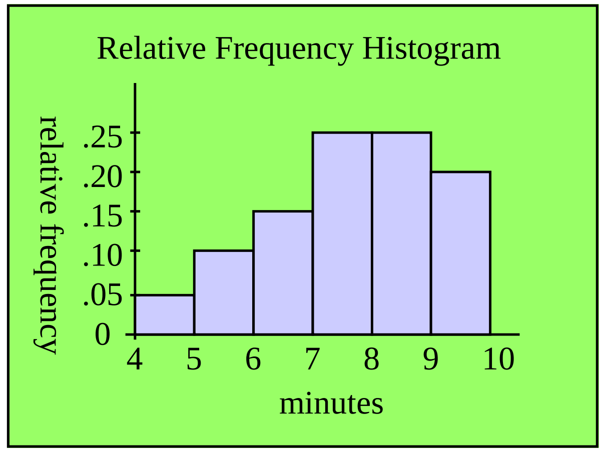
60 people run a mile, record times

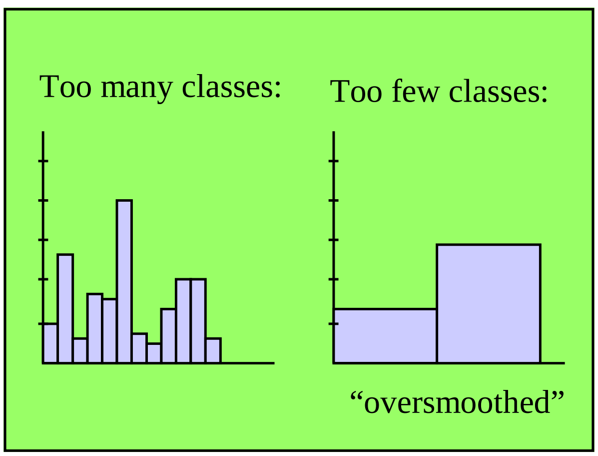
4:55.21 Kevin
….
9:31.07 Greg

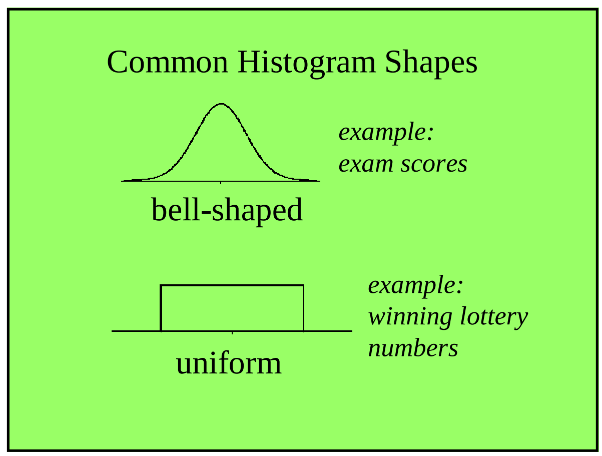
4-5	3
5-6	6
6-7	9
7-8	15
8-9	15
9-10	12
	60

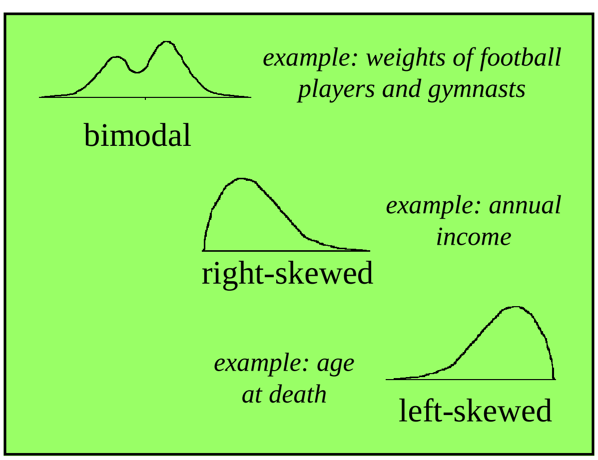
Class	Frequency	Relative Frequency
[4,5)	3	3/60 = .05
[5,6)	6	.10
[6,7)	9	.15
[7,8)	15	.25
[8,9)	15	.25
[9,10)	12	.20
	60	60/60 = 1

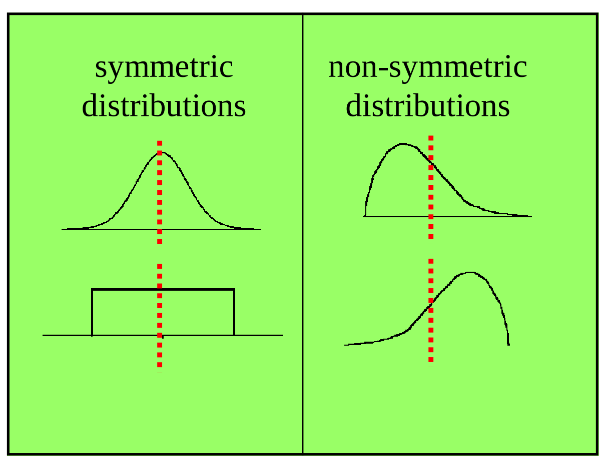

Frequency Histogram
frequency
15
12
9
6
3
0
4 5 6 7 8 9 10
minutes


Relative Frequency Histogram
relative frequency
.25
.20
.15
.10
.05
0
4 5 6 7 8 9 10
minutes

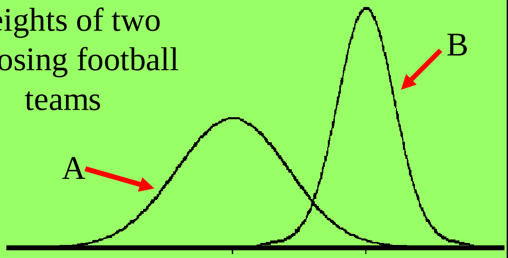

Too many classes:
Too few classes:
“oversmoothed”


Common Histogram Shapes
bell-shaped
example:
exam scores
uniform
example:
winning lottery
numbers


bimodal
example: weights of football
players and gymnasts
right-skewed
example: annual
income
left-skewed
example: age
at death


symmetric
distributions
non-symmetric
distributions

weights of two opposing football teams

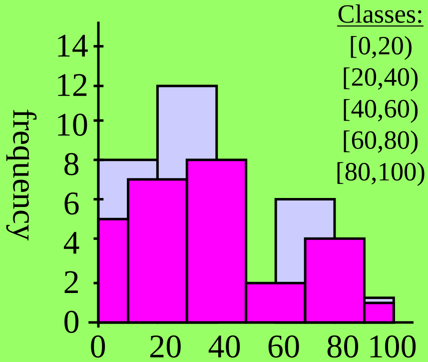


Greater mean? B
Greater variance? A

A downside to histograms:

Choosing different sets of classes can yield different looking histograms for the same data set!

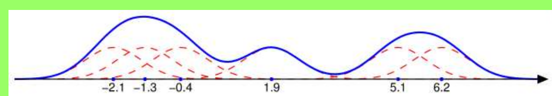
<u>example n = 27</u>
{ 0.3, 0.8, 2.2, 4.7, 5.9, 14.3, 14.4, 15.8, 20.4, 22.2, 22.4, 26.7, 30.2, 33.1, 33.8, 35. 6, 36.4, 36.5, 36.7, 39.2, 64.6, 65.9, 75.7, 76.5, 77.5, 77.8, 93.6 }



Blue <u>Classes:</u>
[0,20)
[20,40)
[40,60)
[60,80)
[80,100)

Pink <u>Classes:</u>
[0,10)
[10,30)
[30,50)
[50,70)
[70,90)
[90,100)

Another option:
Kernel Density Estimation



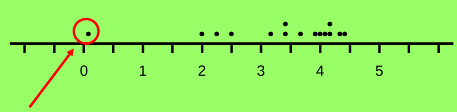
<u>Dot Plot</u>

<u>example</u>

Breaking strength of carbon fibers in GPa (gigapascals)
n=15

{ 0.1, 2.0, 2.3, 2.5, 3.2, 3.4, 3.4, 3.7, 3.9, 4.0, 4.1, 4.2, 4.2, 4.3, 4.4 }

{ 0.1, 2.0, 2.3, 2.5, 3.2, 3.4, 3.4, 3.7, 3.9, 4.0, 4.1, 4.2, 4.2, 4.3, 4.4 }



outlier?

An <u>outlier</u> is an observation that is extremely large or small compared to the majority of the observations.

Stem-and-Leaf Plot

{ 0.1, 2.0, 2.3, 2.5, 3.2, 3.4, 3.4, 3.7, 3.9, 4.0, 4.1, 4.2, 4.2, 4.3, 4.4 }

don't worry about these two columns

stem	leaves	~~number~~	~~depth~~
0	1	~~1~~	~~1~~
1		~~0~~	~~1~~
2	035	~~3~~	~~4~~
3	24479	~~5~~	
4	012234	~~6~~	~~6~~

another option:
0* → 0.0 – 0.4
0• → 0.5 – 0.9
1* → 1.0 – 1.4
etc.

{ 0.1, 2.0, 2.3, 2.5, 3.2, 3.4, 3.4, 3.7, 3.9, 4.0, 4.1, 4.2, 4.2, 4.3, 4.4 }

stem	leaves
0*	1
0•	
1*	
1•	
2*	03
2•	5
3*	244
3•	79
4*	012234
4•	

Yet another option:
0* → 0.0 – 0.1
0t → 0.2 – 0.3
0f → 0.4 – 0.5
0s → 0.6 – 0.7
0• → 0.8 – 0.9
1* → 1.0 – 1.1
etc.

Side-by-side Stem-and-Leaf Plot

More info: of the 15 observations, 7 were from Method A and 8 were from Method B

Method A : n = 7
{ 0.1, 2.0, 2.5, 3.2, 3.4, 3.4, 4.0 }

Method B : n = 8
{ 2.3, 3.7, 3.9, 4.1, 4.2, 4.2, 4.3, 4.4 }

Method A : n = 7
{ 0.1, 2.0, 2.5, 3.2, 3.4, 3.4, 4.0 }

Method B : n = 8
{ 2.3, 3.7, 3.9, 4.1, 4.2, 4.2, 4.3, 4.4 }

stem	A	B
0	1	
1		
2	05	3
3	244	79
4	0	12234

$$\bar{Y}_A = \frac{18.6}{7} = 2.66$$

$$\bar{Y}_B = \frac{31.1}{8} = 3.89$$

n = 4
{ 3.105, 3.114, 3.116, 3.127 }

stem	leaves
3.10	5
3.11	46
3.12	7

possible, but unwise:

stem	leaves
3.1	05 14 16 27

Boxplot

The pth percentile is greater than p% of the data

The 65th percentile is greater than 65% of the data (and less than the other 35% of the data)

Median (50th percentile) = $\tilde{Y}$

1 – rank data from smallest to largest

2 – if n is odd, median = middle value
if n is even, median = mean of the two middle values

Book notation: ordered data set

$$\{Y_{(1)}, Y_{(2)}, \ldots Y_{(n)}\}$$

location of the median = $l_m = \dfrac{n+1}{2}$

if n is odd → median = $Y_{(l_m)}$

if n is even → median = $\dfrac{Y_{(l_m - \frac{1}{2})} + Y_{(l_m + \frac{1}{2})}}{2}$

Four ranked data sets

n	6	7	5681	6004
l_m	$\frac{6+1}{2} = 3.5$	$\frac{7+1}{2} = 4$	$\frac{5681+1}{2} = 2841$	$\frac{6004+1}{2} = 3002.5$
median	mean of 3rd & 4th values	4th value	2841st value	mean of 3002nd & 3003rd values

X X X X X X

X X X X X X X

Quartiles

Q_1 = first/lower quartile (25th percentile)

$\tilde{Y}$ = Q_2 = median = second/middle quartile

Q_3 = third/upper quartile (75th percentile)

Q_1 = median of the lower half of the data (include the median if n is odd)

Q_3 = median of the upper half of the data (include the median if n is odd)

Book notation:

location of the quartiles =

$l_q = \dfrac{n+3}{4}$ if n is odd

$l_q = \dfrac{n+2}{4}$ if n is even

$Q_1 = Y_{(l_q)}$

$Q_3 = Y_{(n+1-l_q)}$

n	6	7	5681	6004
l_q	$\frac{6+2}{4}=2$	$\frac{7+3}{4}=2.5$	$\frac{5681+3}{4}=1421$	$\frac{6004+2}{4}=1501.5$
Q_1	2nd value	mean of 2nd & 3rd values	1421st value	mean of 1501st & 1502nd values
$n+1-l_q$	$6+1-2$ $=5$	$7+1-2.5$ $=5.5$	$5681+1-1421$ $=4261$	$6004+1-1501.5$ $=4503.5$
Q_3	5th value	mean of 5th & 6th values	4261st value	mean of 4503rd & 4504th values

n = 7 { 0, 11, 13, 25, 34, 40, 81 }

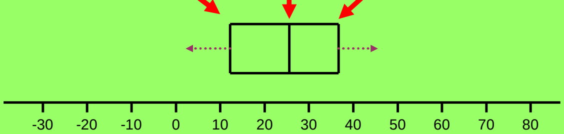
$Q_1 = \frac{11+13}{2} = 12$ median = 25

$Q_3 = \frac{34+40}{2} = 37$

range = max – min = 81 – 0 = 81

interquartile range = IQR = $Q_3 - Q_1$ =
37 – 12 = 25

Q_1 = 12 median = 25 Q_3 = 37



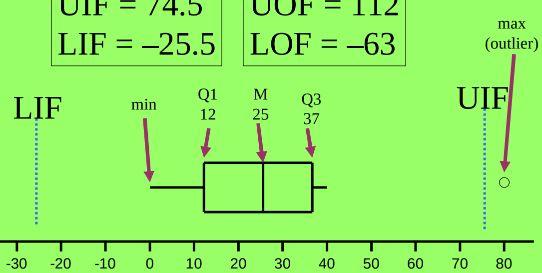
The whiskers are going to try to reach out to the min and max but may not be able to!

step = 1.5(IQR) = 1.5(25) = 37.5

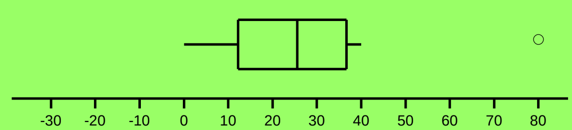
upper inner fence = UIF = Q_3 + step = 37 + 37.5 = 74.5	lower inner fence = LIF = Q_1 – step = 12 – 37.5 = –25.5
upper outer fence = UOF = Q_3 + 2(step) = 37 + 2(37.5) = 112	lower outer fence = LOF = Q_1 – 2(step) = 12 – 2(37.5) = –63

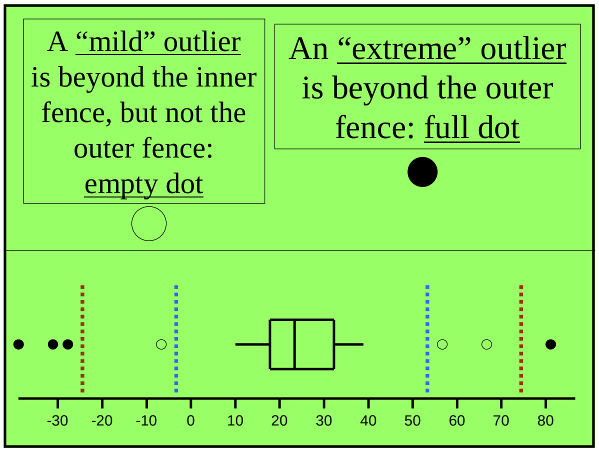
UIF = 74.5
LIF = –25.5

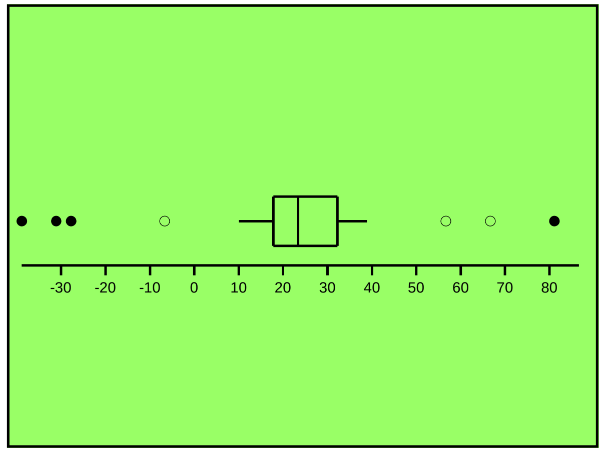
UOF = 112
LOF = –63

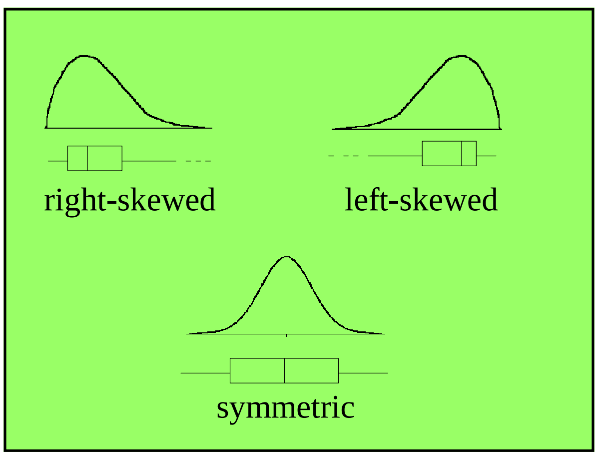


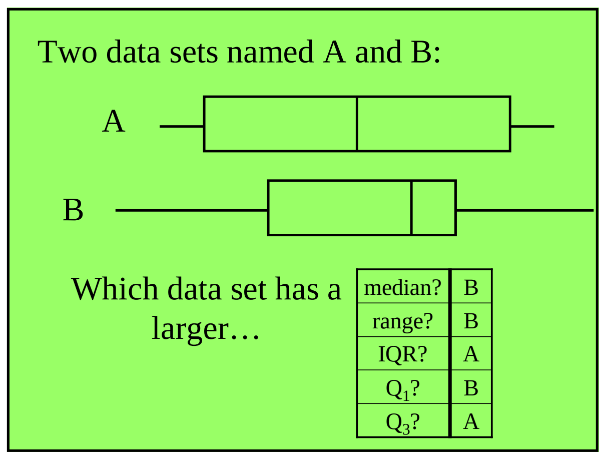
{ 0, 11, 13, 25, 34, 40, 81 }




A "mild" outlier is beyond the inner fence, but not the outer fence: empty dot
An "extreme" outlier is beyond the outer fence: full dot
-30
-20
-10
0
10
20
30
40
50
60
70
80


-30
-20
-10
0
10
20
30
40
50
60
70
80


right-skewed
left-skewed
symmetric


Two data sets named A and B:
A
B
Which data set has a larger…

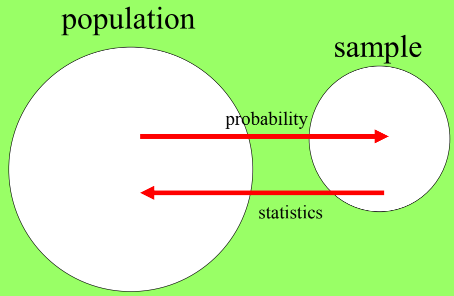
median?	B
range?	B
IQR?	A
Q_1?	B
Q_3?	A

MA 3710

3.1 – 3.4

Pyramids and Sphinx
Giza, Egypt

Luxor Casino
Las Vegas



set – an unordered collection of zero or more elements

alphabet = { a, b, … z } = { q, x, … c }

positive integers = { 1, 2, 3, … }

people over 200 years old = { } = $\varnothing$

empty set = { } = $\varnothing$

the set of zero elements

subset – a set where each element is also a member of another set

$A \subseteq B$ if every element in A is also in B

vowels = {a, e, i, o, u } $\subseteq$ alphabet = { a, b, … z }

even positive integers $\subseteq$ positive integers
{ 2, 4, 6, … } $\subseteq$ { 1, 2, 3, … }

$A \subseteq A$ every set is a subset of itself

$\varnothing \subseteq A$ the empty set is a subset of all sets

how many subsets of { a, b, … z } exist?

{ a, b, c, … z }

$2 \times 2 \times 2 \times \ldots \times 2 = 2^{26} = 67{,}108{,}864$

if a set has K elements, then there are 2^K subsets

example outcomes of a coin flip

{ H, T } $\rightarrow$ 2 elements $\rightarrow$ $2^2 = 4$ subsets

subsets:

$\varnothing$ { H } { T } { H, T }

experiment - a process that has variation in its outcomes

sample space = S = the set of all possible outcomes of an experiment

event - a subset of the sample space

probability of an event A = P(A)

a value between 0 and 1 representing how likely it is and the proportion of time it will occur in the long run

$\boxed{0 \leq P(A) \leq 1}$

if P(A) = 1, it will happen P(S) = 1

if P(A) = 0, it won't happen P(∅) = 0

$$P(A) = \frac{\text{“size” of event A}}{\text{“size” of S}}$$

the long-run relative frequency of an event approaches its probability as the sample size gets large

$$\frac{\text{\# of occurences of A}}{\text{\# of experiments}} \xrightarrow{\text{large n}} P(A)$$

if P(A) is unknown, it can be estimated by repeating the experiment many times and observing how often it occurs

Computer simulation of coin flips

$$\frac{1}{10} = .10$$

$$\frac{45}{100} = .45$$

$$\frac{471}{1000} = .471$$

$$\frac{5047}{10000} = .5047$$

$$\frac{50064}{100000} = .50064$$

example roll a die once

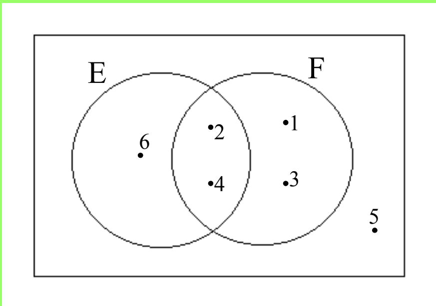
S = { 1, 2, 3, 4, 5, 6 }

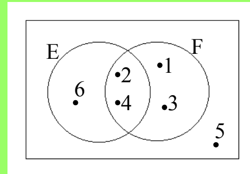
There are 2^6 = 64 different events

E = { 2, 4, 6 } = an even number appears
P(E) = 3/6 = 1/2

F = { 1, 2, 3, 4 } = a number less than or equal to four appears
P(F) = 4/6 = 2/3

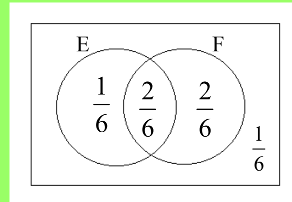
Venn Diagram



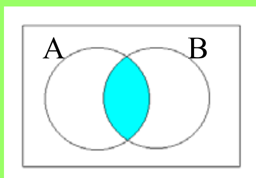


E = { 2, 4, 6 } = an even number appears
P(E) = 3/6 = 1/2

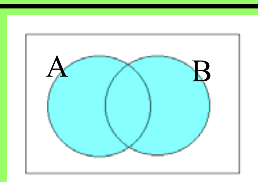
F = { 1, 2, 3, 4 } = a number less than or equal to four appears
P(F) = 4/6 = 2/3



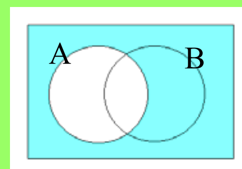
<u>intersection</u>
$A \cap B$
"A and B"

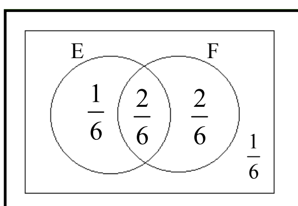


<u>union</u>
$A \cup B$
"A or B or both"



<u>complement</u>
$\overline{A}$
"not A"





die example continued
E = an even number appears
F = a number less than or equal to four appears

$P(E \cap F) = \frac{2}{6} = \frac{1}{3}$ 

$P(\overline{E}) = \frac{3}{6} = \frac{1}{2}$ 

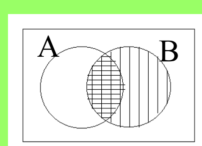
$P(E \cup F) = \frac{5}{6}$ 

$P(\overline{F}) = \frac{2}{6} = \frac{1}{3}$ 

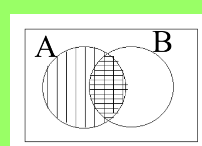
conditional probability

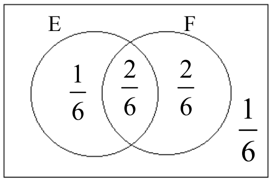
probability of A given B = $P(A|B)$

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



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





die example continued
E = an even number appears
F = a number less than or equal to four appears

$$P(E\,|\,F) = \frac{P(E \cap F)}{P(F)} = \frac{2/6}{4/6} = \frac{2}{4} = \frac{1}{2}$$

Notice that in this example, $P(E\,|\,F) = P(E) = \frac{1}{2}$

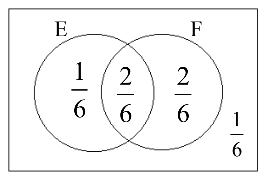
We call E and F <u>independent</u> events.

A and B are <u>independent</u> if the occurrence of one event doesn't affect the probability of the other.

A and B are <u>independent</u> iff $\boxed{\begin{array}{l} P(A\,|\,B) = P(A) \\ P(B\,|\,A) = P(B) \\ P(A \cap B) = P(A) \times P(B) \end{array}}$

A and B are <u>mutually exclusive</u> if it is impossible for both to occur simultaneously.

A and B are <u>mutually exclusive</u> iff $\boxed{P(A \cap B) = 0}$



die example continued
E = an even number appears
F = a number less than or equal to four appears

$P(E \cap F) = \frac{2}{6} \neq 0 \rightarrow$ E and F are <u>not</u> mutually exclusive

$P(G \cap H) = 0 \rightarrow$ G and H are mutually exclusive

G = { 1, 2 } = a number less than three appears

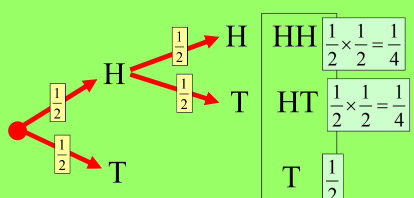
H = { 4, 5, 6 } = a number greater than three appears

<u>example</u>

toss a coin

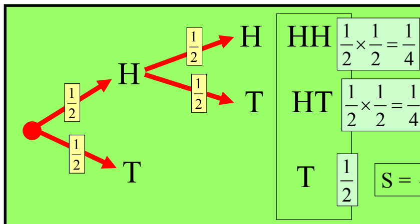
if H $\longrightarrow$ toss once more

if T $\longrightarrow$ stop



<u>Tree diagram</u>

S = { HH, HT, T }



S = { HH, HT, T }

P(H appears at some point in the experiment | T appears at some point in the experiment) =

$$\text{P(H appears | T appears)} = \frac{\text{P(H appears} \cap \text{T appears)}}{\text{P(T appears)}} = \frac{1/4}{3/4} = \frac{1}{3}$$

Multiplicative laws

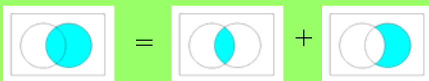
$P(A \cap B) = P(A) \times P(B \mid A)$

$P(A \cap B) = P(B) \times P(A \mid B)$

If A and B are independent $\rightarrow$ $P(A \cap B) = P(A) \times P(B)$

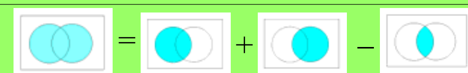
Law of total probability

$$P(B) = P(A \cap B) + P(\overline{A} \cap B)$$



Additive Law

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



Complement Rule

$$P(\overline{A}) = 1 - P(A)$$

A company makes fluorescent light bulbs in three factories: Rockford, Sycamore, and Thompson. One bulb is selected at random.

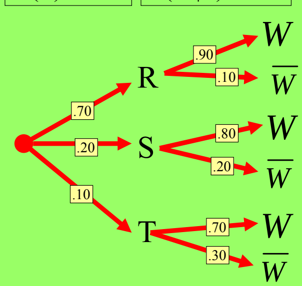
70% are made in Rockford: P(R) = .70
20% are made in Sycamore: P(S) = .20
10% are made in Thompson: P(T) = .10

W = a bulb is still working after 10,000 hours of use

Rockford has a 90% success rate: P(W|R) = .90
Sycamore has a 80% success rate: P(W|S) = .80
Thompson has a 70% success rate: P(W|T) = .70

P(R) = .70
P(S) = .20
P(T) = .10

P(W|R) = .90
P(W|S) = .80
P(W|T) = .70



RW $.70 \times .90 = .63$

$R\overline{W}$ $.70 \times .10 = .07$

SW $.20 \times .80 = .16$

$S\overline{W}$ $.20 \times .20 = .04$

TW $.10 \times .70 = .07$

$T\overline{W}$ $.10 \times .30 = .03$

RW	.63
$R\overline{W}$	.07
SW	.16
$S\overline{W}$	.04
TW	.07
$T\overline{W}$	.03

P(W) = .63 + .16 + .07 = .86

$P(R \cap W)$ $P(S \cap W)$ $P(T \cap W)$

law of total probability

$P(S \cup \overline{W}) = .07 + .16 + .04 + .03 = .30$

could have used additive law:

$P(S \cup \overline{W}) = P(S) + P(\overline{W}) - P(S \cap \overline{W})$

$.20 + .14 - .04 = .30$

$P(R \cap W)$

Probability Table

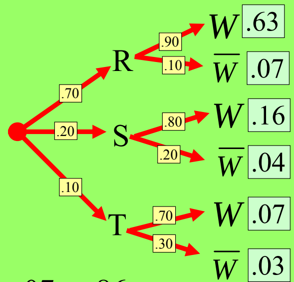
P(R) = .70
P(S) = .20
P(T) = .10

	W	$\overline{W}$	
R	.63	.07	.70
S	.16	.04	.20
T	.07	.03	.10
	.86	.14	1

.90

$$P(W \mid R) = \frac{P(R \cap W)}{P(R)} \quad \begin{matrix} .63 \\ .70 \end{matrix}$$

	W	$\overline{W}$	
R	.63	.07	.70
S	.16	.04	.20
T	.07	.03	.10
	.86	.14	1



P(W) = .63 + .16 + .07 = .86

$P(S \cup \overline{W}) = .07 + .16 + .04 + .03 = .30$

Bayes Rule

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B \mid A)}{\underbrace{P(A) \times P(B \mid A)}_{P(A \cap B)} + \underbrace{P(\overline{A}) \times P(B \mid \overline{A})}_{P(\overline{A} \cap B)}}$$

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(A) \times P(B \mid A) + P(\overline{A}) \times P(B \mid \overline{A})}$$

1% of the people on an island have Disease X. A test for the disease is 99% accurate. Given that someone on the island tests positive, what is the probability they actually have the disease?

Given:

$P(D) = .01$

$P(+ \mid D) = .99$

$P(- \mid \overline{D}) = .99$

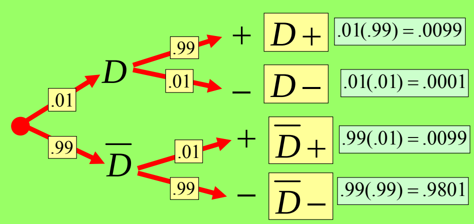
Asked:

$$P(D \mid +) = ?$$

$P(D) = .01$

$P(+ \mid D) = .99$

$P(- \mid \overline{D}) = .99$



$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(A) \times P(B \mid A) + P(\overline{A}) \times P(B \mid \overline{A})}$$

$$P(D \mid +) = \frac{P(D) \times P(+ \mid D)}{P(D) \times P(+ \mid D) + P(\overline{D}) \times P(+ \mid \overline{D})}$$

$$P(D \mid +) = \frac{.01(.99)}{.01(.99) + .99(.01)} = \frac{.0099}{.0099 + .0099} = \frac{1}{2}$$

The probability is only .50 ??!???

Plan: if a person tests positive, try again!

Given that someone tests positive *three times in a row*, what is the probability they actually have the disease?

$$P(D \mid +++) = \frac{P(D) \times P(+++ \mid D)}{P(D) \times P(+++ \mid D) + P(\overline{D}) \times P(+++ \mid \overline{D})}$$

$$P(D \mid +++) = \frac{.01(.99)^3}{.01(.99)^3 + .99(.01)^3} = .99989798$$

Coin flips

Four flips: $P(HTTH) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$

Five flips: P(at least one T) = 1 – P(no T)

$$= 1 - P(HHHHH) = 1 - \left(\frac{1}{2}\right)^5 = 1 - \frac{1}{32} = \frac{31}{32}$$

Ten flips:

Which outcome is most likely?:

HHHHHHHHHH
TTTTTTTTTT
THHTHTTHHT

All three have the same probability!

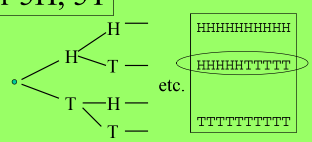
$$\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}$$

Ten flips:

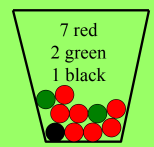
P(5 H, 5 T in any order) = ?

$$= (\quad)\left(\frac{1}{2}\right)^{10} = 252\left(\frac{1}{2}\right)^{10} = .246$$

# of ways to order 5H, 5T



A basket contains 10 balls:



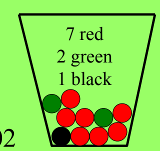
3 will be drawn out one at a time at random

With or without replacement?

Independent draws

Dependent draws

With replacement:



$P(RRR) = (.70)^3 = .343$

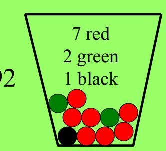
$P(BGB) = (.10)(.20)(.10) = .002$

P(no black draws) = $(.9)^3 = .729$

P(at least one green) = 1 – P(no G) =

$$= 1 - (.8)^3 = .488$$

Without replacement:



$P(RRR) = \left(\frac{7}{10}\right)\left(\frac{6}{9}\right)\left(\frac{5}{8}\right) = .292$

$P(BGB) = \left(\frac{1}{10}\right)\left(\frac{2}{9}\right)\left(\frac{0}{8}\right) = 0$

P(no black draws) = $\left(\frac{9}{10}\right)\left(\frac{8}{9}\right)\left(\frac{7}{8}\right) = .700$

P(at least one green) = 1 – P(no G) =

$$= 1 - \left[\left(\frac{8}{10}\right)\left(\frac{7}{9}\right)\left(\frac{6}{8}\right)\right] = .533$$

The famous birthday problem:

Would you be surprised if two people in this room shared the same birthday?

P(at least one shared birthday) =
= 1 – P(no shared birthdays) =

<u>assume</u>: independent birthdays
equally likely birthdays

P(at least one shared birthday) =
= 1 – P(no shared birthdays) =

$$1-\left[\underbrace{\left(\frac{365}{365}\right)}_{\text{Person 1}}\underbrace{\left(\frac{364}{365}\right)}_{\text{Person 2}}\underbrace{\left(\frac{363}{365}\right)}_{\text{Person 3}}\cdots\underbrace{\left(\frac{365-(n-1)}{365}\right)}_{\text{Person n}}\right]=$$

n	P(at least one match)
10	.1169
23	.5073
32	.7533
50	.9704

A <u>random variable</u> Y assigns a value to every outcome in the sample space

<u>Now</u>: a <u>discrete random variable</u> Y has a countable (*listable & not "squeezable"*) set of possible values

<u>Later</u>: a <u>continuous random variable</u> Y has an uncountable (*unlistable &"squeezable"*) set of possible values

Y = random variable y = a constant

A = # of heads in three coin flips
A = { 0, 1, 2, 3 }
$\boxed{\text{discrete}}$

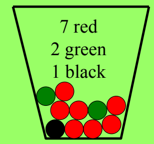
B = height in cm
~~B = { 0,~~
$\boxed{\text{continuous}}$

C = # of heads before first tail
C = { 0, 1, 2, 3, 4, 5, … }
$\boxed{\text{discrete}}$

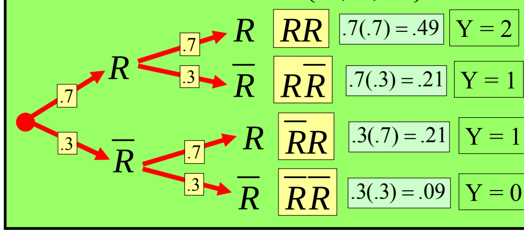
D = time to run one mile
$\boxed{\text{continuous}}$

<u>Disclaimer</u>: There exist countable sets that are "squeezable".

Draw two balls out one at a time at random with replacement



Y = # of red draws = { 0, 1, 2 }



Every discrete random variable has a probability function p(y) = P(Y = y)

$$0 \le p(y) \le 1$$

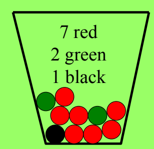
$$\sum_y p(y) = 1$$

Every discrete random variable has a cumulative distribution function (cdf)

$$F(y) = P(Y \le y)$$

Y = # of red draws = { 0, 1, 2 }

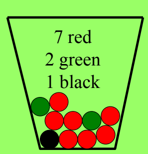
RR	.7(.7) = .49	Y = 2
$R\overline{R}$	.7(.3) = .21	Y = 1
$\overline{R}R$	.3(.7) = .21	Y = 1
$\overline{R}\,\overline{R}$	.3(.3) = .09	Y = 0



y	P(Y=y)	F(y)
0	.09	.09
1	.42	.51
2	.49	1

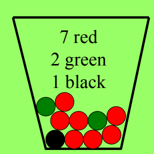
1

Same basket, different variable: now draw repeatedly until black appears



Y = # of draws until black appears = {1, 2, 3, 4, 5, … }

y	P(Y=y)	what would we observe?
1	.1	B
2	(.9)(.1)	$\overline{B}B$
3	$(.9)^2(.1)$	$\overline{B}\,\overline{B}B$
…		
y	$(.9)^{y-1}(.1)$	$\overline{B}\,\overline{B}...\overline{B}B$



Y = # of draws until black appears = {1, 2, … }

geometric distribution

$$p(y) = (.9)^{y-1}(.1) \quad \text{if } y = 1, 2, 3, \ldots$$
$$= 0 \quad \text{otherwise}$$

$$\sum_{y=1}^{\infty} (.9)^{y-1}(.1) \to 1$$

Population mean (expected value) of a discrete random variable

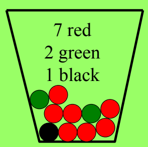
$\bar{Y}$ = sample mean $\quad \boxed{\mu = \text{population mean}}$

$$\mu = E(Y) = \Sigma\, y \cdot P(Y = y)$$

$\bar{Y} \to \mu$ as n gets large

Y = # of red draws = { 0, 1, 2 }

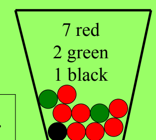
y	P(Y=y)
0	.09
1	.42
2	.49



$$\mu = E(Y) = 0(.09) + 1(.42) + 2(.49) = 1.4$$

After many experiments, the average number of red balls observed will be close to 1.4

Y = # of draws until black appears = {1, 2, 3, 4, 5, … }



$$p(y) = (.9)^{y-1}(.1) \quad \text{if } y = 1, 2, 3, \ldots$$
$$= 0 \quad \text{otherwise}$$

$$\mu = E(Y) = \sum_{y=1}^{\infty} y(.9)^{y-1}(.1) \to 10$$

On average, we wait 10 draws to see the next black ball.

Four buses carrying 40, 33, 25, and 50 students. (148 total)

*select one student at random
Y_1 = # of students on bus he/she rode

*select one bus driver at random
Y_2 = # of students on bus he/she drove

$$E(Y_1) \quad \begin{array}{|c|} > \\ < \\ = \\ \hline \end{array} \quad E(Y_2)$$

Four buses carrying 40, 33, 25, and 50 students. (148 total)
*select one student at random
Y1 = # of students on bus he/she rode
*select one bus driver at random
Y2 = # of students on bus he/she drove

$$E(Y_1) = 40\left(\frac{40}{148}\right) + 33\left(\frac{33}{148}\right) + 25\left(\frac{25}{148}\right) + 50\left(\frac{50}{148}\right) = 39.28$$

$$E(Y_2) = 40\left(\frac{1}{4}\right) + 33\left(\frac{1}{4}\right) + 25\left(\frac{1}{4}\right) + 50\left(\frac{1}{4}\right) = 37$$

$$\boxed{E(Y_1) > E(Y_2)}$$

example roulette

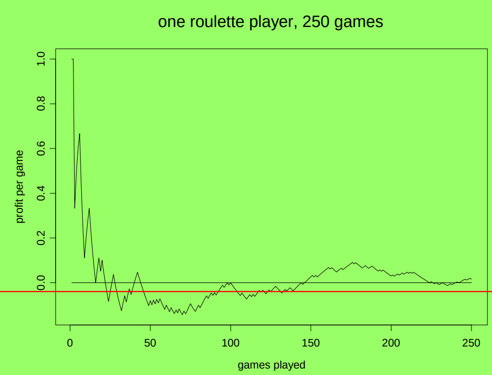
1 – 36 (18 reds, 18 blacks)
0, 00 (2 greens)

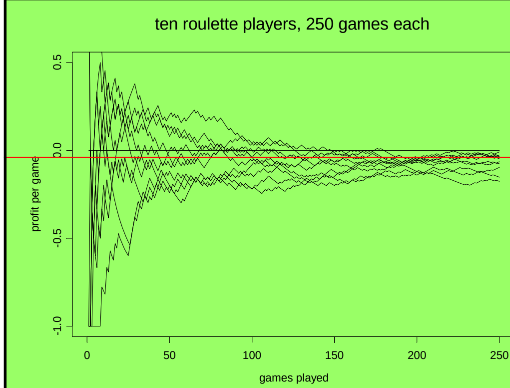
Y = profit from a \$1 bet on black = { –1, 1 }

y	P(Y=y)
–1	$^{20}\!/_{38}$
1	$^{18}\!/_{38}$

$$\mu = E(Y) = -1\left(\frac{20}{38}\right) + 1\left(\frac{18}{38}\right) = -.0526$$

$\boxed{-5.26\ ¢}$





Doubling-Up Strategy

Goal: win \$1

Bet \$1.	If you win, stop.	If you lose,…
Bet \$2.	If you win, stop.	If you lose,…
Bet \$4.	If you win, stop.	If you lose,…
Bet \$8.	If you win, stop.	If you lose,…

the catch: run out of money
maximum table bets

variance and standard deviation
of a discrete random variable

s^2 = sample variance
s = sample standard deviation

σ^2 = population variance
σ = population standard deviation

$s^2 \to \sigma^2$
$s \to \sigma$ as n gets large

$$\sigma^2 = Var(Y) = E\left[(Y-\mu)^2\right]$$

$$\sigma^2 = \underbrace{\sum_y (y-\mu)^2 \times p(y)}_{\text{"regular"}} = \underbrace{E(Y^2) - \mu^2}_{\text{"convenient"}}$$

$$\sigma = SD(Y) = \sqrt{\sigma^2}$$

Y = outcome of a loaded die = { 1, 2, 3, 4, 5, 6 }

Y	P(Y= y)	y–μ	(y–μ)^2	y^2
1	.1	–3	9	1
2	.1	–2	4	4
3	.2	–1	1	9
4	.2	0	0	16
5	.1	1	1	25
6	.3	2	4	36

first find μ:

$$\mu = E(Y) = 1(.1) + 2(.1) + 3(.2) + 4(.2) + 5(.1) + 6(.3) = 4$$

Y	P(Y= y)	y–μ	(y–μ)^2	y^2
1	.1	–3	9	1
2	.1	–2	4	4
3	.2	–1	1	9
4	.2	0	0	16
5	.1	1	1	25
6	.3	2	4	36

$$\sigma^2 = \sum_y (y-\mu)^2 \times p(y) = $$
$$= \begin{bmatrix} 9(.1) + 4(.1) + 1(.2) + \\ 0(.2) + 1(.1) + 4(.3) \end{bmatrix} = \boxed{2.8}$$

Y	P(Y= y)	y–μ	(y–μ)^2	y^2
1	.1	–3	9	1
2	.1	–2	4	4
3	.2	–1	1	9
4	.2	0	0	16
5	.1	1	1	25
6	.3	2	4	36

$$\sigma^2 = E(Y^2) - \mu^2$$
$$= \begin{bmatrix} 1(.1) + 4(.1) + 9(.2) + \\ +16(.2) + 25(.1) + 36(.3) \end{bmatrix} - (4)^2 = \boxed{2.8}$$

$$\sigma^2 = \sum_y (y-\mu)^2 \times p(y) = $$
$$= \begin{bmatrix} 9(.1) + 4(.1) + 1(.2) + \\ 0(.2) + 1(.1) + 4(.3) \end{bmatrix} = 2.8$$

$$\sigma^2 = E(Y^2) - \mu^2$$
$$= \begin{bmatrix} 1(.1) + 4(.1) + 9(.2) + \\ +16(.2) + 25(.1) + 36(.3) \end{bmatrix} - (4)^2 = 2.8$$

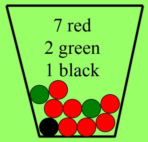
$$\sigma = \sqrt{\sigma^2} = \sqrt{2.8} = \boxed{1.67}$$

Computer Simulation: the loaded die was rolled 1000 times.

$$n = 1000$$
$$\overline{Y} = 4.043 \approx \mu = 4$$
$$s^2 = 2.84 \approx \sigma^2 = 2.8$$
$$s = 1.69 \approx \sigma = 1.67$$

Y = # of red draws = { 0, 1, 2 }

y	P(Y=y)
0	.09
1	.42
2	.49



$$\mu = E(Y) = 0(.09) + 1(.42) + 2(.49) = 1.4$$

From previous work, we know that $\mu = 1.4$

y	P(Y=y)	y^2
0	.09	0
1	.42	1
2	.49	4

$$\sigma^2 = Var(Y) = E(Y^2) - \mu^2$$
$$= \left[0(.09) + 1(.42) + 4(.49)\right] - (1.4)^2 = 0.42$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.42} = .648$$

Combination/Choosing Symbol, a.k.a. the Binomial Coefficient

How many options do we have if we want to choose 2 people from 4 (unordered, without replacement)?

{ Al, Bob, Carla, Deb }

{ Al, Bob } {Bob, Carla }
{ Al, Carla } {Bob, Deb }
{ Al, Deb } {Carla, Deb }

$\boxed{6}$

$\binom{n}{y}$ = " n choose y" = # of ways to choose an unordered group of y objects from a group of n without replacement

$$\boxed{\binom{n}{y} = \frac{n!}{y!(n-y)!}}$$

$$\binom{4}{2} = \frac{4!}{2!2!} = \frac{4\times3\times2\times1}{(2\times1)(2\times1)} = \frac{24}{4} = 6$$

50 people, choose 7 to win $10

$$\binom{50}{7} = \frac{50!}{7!43!} = 99,884,400$$

$$\binom{50}{43} = \frac{50!}{43!7!} = 99,884,400$$

$$\binom{50}{1} = \frac{50!}{1!49!} = 50$$

$$\binom{50}{0} = \frac{50!}{0!50!} = 1$$

$$\boxed{\binom{n}{y} = \binom{n}{n-y}}$$

$$\boxed{\binom{n}{1} = n}$$

$$\boxed{\binom{n}{0} = 1}$$

Binomial Random Variable
a special discrete random variable

Y = # of successes in n trials

Y = { 0, 1, … n }

<u>The following conditions must hold</u>:

n must be fixed

each trial has exactly two possible outcomes: success or failure

P(success) = p and P(failure) = 1 – p = q do not change from trial to trial

trials must be independent

<u>Are these random variables Binomial?</u>

A = # of tails in 6 coin flips
A = { 0, 1, 2, 3, 4, 5, 6 }

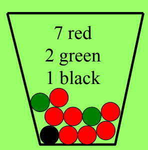
Yes: *n = 6, p = .50*

B = # flips needed to have a total of 3 tails
B = { 3, 4, 5, … }

No: *n isn't fixed*

<u>Are these random variables Binomial?</u>

C = # of red draws if we draw 4 balls out of this basket at random without replacement



C = { ~~0~~, 1, 2, 3, 4 }

No: *p changes, and trials are dependent*

D = same as above, but with replacement

D = { 0, 1, 2, 3, 4 }

Yes: *n = 4, p = .70*

A company makes fluorescent light bulbs

W = a bulb is working after 10,000 hours of use

P(W) = .86 assume bulbs are independent

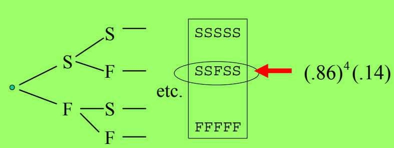
In a batch of n = 5 bulbs, what is the probability exactly 4 will last 10,000 hours?

Y = # of successful bulbs

Y is Binomial with n = 5, p = .86

Y is Binomial with n = 5, p = .86

$$P(Y=4)=\binom{5}{4}(.86)^4(.14)=.3829$$



Y is Binomial with n = 5, p = .86

$$P(Y=0)=\binom{5}{0}(.86)^0(.14)^5=.000054$$

$$P(Y=1)=\binom{5}{1}(.86)^1(.14)^4=.0017$$

$$P(Y=2)=\binom{5}{2}(.86)^2(.14)^3=.0203$$

$$P(Y=3)=\binom{5}{3}(.86)^3(.14)^2=.1247$$

$$P(Y=4)=\binom{5}{4}(.86)^4(.14)^1=.3829$$

$$P(Y=5)=\binom{5}{5}(.86)^5(.14)^0=.4704$$

Y is Binomial with n = 20, p = .70

$$P(Y=13)=\binom{20}{13}(.70)^{13}(.30)^{7}=.1643$$

In general: n trials, P(success) = p, P(failure) = q

$$P(y)=\begin{cases}\binom{n}{y}p^{y}q^{n-y} & y=0,1,\ldots n\\ 0 & \text{otherwise}\end{cases}$$

Remember this problem?

Ten flips of a coin:

P(5 H, 5 T in any order) $= 252\left(\frac{1}{2}\right)^{10} = .246$

Y = # heads in 10 flips

heads ~ Binomial(n=10, p=.5)

$$P(Y=5)=\binom{10}{5}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{5}=.246$$

$$\mu = E(Y) = \Sigma\, y\cdot P(Y=y) = np$$

$$\sigma^2 = Var(Y) = E(Y^2)-\mu^2 = npq$$

$$\sigma = SD(Y) = \sqrt{\sigma^2} = \sqrt{npq}$$

any discrete r.v. | binomial only

Basketball free throws:
n = 20, p = .70

$$\mu = E(Y) = np = 20(.70) = 14$$

$$\sigma^2 = Var(Y) = npq = 20(.70)(.30) = 4.2$$

$$\sigma = SD(Y) = \sqrt{npq} = \sqrt{4.2} = 2.05$$

Batches of n = 5 light bulbs, p = .86

$$\mu = E(Y) = np = 5(.86) = 4.3$$

$$\sigma^2 = Var(Y) = npq = 5(.86)(.14) = .602$$

$$\sigma = SD(Y) = \sqrt{npq} = \sqrt{.602} = .776$$

A class has 9 women and 6 men. If the class meets 10 times with perfect attendance at each meeting, and a student is selected at random at each meeting, what is

P(a woman is selected 7 times) =

women ~ Binomial(n=10, p=.6)

$$P(Y=7)=\binom{10}{7}(.6)^{7}(.4)^{3}=.215$$

P(a man is selected 3 times)=

men ~ Binomial(n=10, p=.4)

$$P(Y=3)=\binom{10}{3}(.4)^{3}(.6)^{7}=.215$$

Geometric Random Variable
a special discrete random variable

*closely related to Binomial (same rules but n isn't fixed)

Y = # of trials needed to observe first success

Y = { 1, 2, … }

P(success) = p

$$P(y) = \begin{cases} (1-p)^{y-1}p & y = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

If Y is Geometric, then

$$\mu = E(Y) = \frac{1}{p}$$

$$\sigma^2 = Var(Y) = \frac{1-p}{p^2}$$

$$\sigma = SD(Y) = \sqrt{\frac{1-p}{p^2}} = \frac{\sqrt{1-p}}{p}$$

A new radio communcations system attempts to send a signal to a receiver in the basement of the EERC once a minute. The probability of success on each attempt (independently) is .25.

Y = # of attempts needed to observe first success

$$P(Y=3) = P(FFS) =$$

$$(1-.25)^{3-1}(.25) =$$

$$= (.75)^2(.25) = .1406$$

A new radio commincations system attempts to send a signal to a receiver in the basement of the EERC once a minute. The probability of success on each attempt (independently) is .25.

Y = # of attempts needed to observe first success

On average, the # of attempts needed is

$$\mu = E(Y) = \frac{1}{p} = \frac{1}{.25} = 4$$

$$\sigma^2 = Var(Y) = \frac{1-p}{p^2} = \frac{1-.25}{(.25)^2} = 12$$

$$\sigma = \sqrt{12} = 3.46$$

Poisson Random Variable
a special discrete random variable

Y = # of occurrences of an event in a specified time period

Y = { 0, 1, 2, 3, … }

λ = mean rate of occurrences

$$\lambda > 0$$

$$P(y) = \begin{cases} \dfrac{\lambda^y}{y!}e^{-\lambda} & y = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{y=o}^{\infty} \frac{\lambda^y}{y!}e^{-\lambda} = e^{-\lambda}\sum_{y=o}^{\infty}\frac{\lambda^y}{y!} = e^{-\lambda}e^{\lambda} = 1$$

If Y is Poisson, then

$$\mu = E(Y) = \lambda$$
$$\sigma^2 = Var(Y) = \lambda$$
$$\sigma = SD(Y) = \sqrt{\lambda}$$

Y = # of lightning strikes to ME-EM per year

Y is Poisson with $\lambda = 0.125 = \frac{1}{8}$

$$\mu = E(Y) = \lambda = \frac{1}{8}$$

On average, we expect $\frac{1}{8}$ of a lightning strike per year (one lightning strike about every 8 years).

$$P(Y) = \frac{\left(\frac{1}{8}\right)^y}{y!} e^{-\frac{1}{8}} \quad \text{if } y = 0, 1, \ldots$$

P(no strikes for the next year) =

$$P(Y=0) = \frac{\left(\frac{1}{8}\right)^0}{0!} e^{-\frac{1}{8}} = e^{-\frac{1}{8}} = .88250$$

P(at least one strike within the next year) =
$P(Y \geq 1) = 1 - P(Y=0) = 1 - .88250 = .11750$

P(exactly 3 strikes within the next year) =

$$P(Y=3) = \frac{\left(\frac{1}{8}\right)^3}{3!} e^{-\frac{1}{8}} = .000287 \approx \frac{1}{3481}$$

Y = # of orders placed per day on a sales website

Y is Poisson with $\lambda = 5$

$$\mu = E(Y) = \lambda = 5$$
$$\sigma^2 = Var(Y) = \lambda = 5$$
$$\sigma = SD(Y) = \sqrt{\lambda} = \sqrt{5} = 2.236$$

P(three or more orders will be placed tomorrow) =

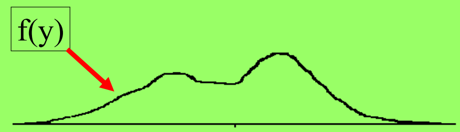
$$P(Y \geq 3) = 1 - P(Y \leq 2) =$$
$$= 1 - \left[P(Y=0) + P(Y=1) + P(Y=2)\right] =$$
$$1 - \left[\frac{5^0}{0!}e^{-5} + \frac{5^1}{1!}e^{-5} + \frac{5^2}{2!}e^{-5}\right] = 1 - e^{-5}\left[1 + 5 + \frac{25}{2}\right] =$$
$$= 1 - .12465 = .87535$$

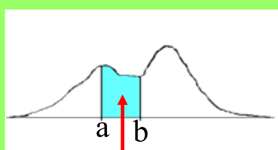
Two other discrete r.v.'s found in the text:

Negative Binomial
Hypergeometric

A <u>continuous random variable</u> Y has an uncountable (*unlistable & "squeezable"*) set of possible values <u>over an interval</u>

f(y) = probability density function





$P(a < Y < b) =$ the area under f(y) between a and b

$$f(y) \geq 0$$

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

$$F(y_0) = P(Y \leq y_0) = \int_{-\infty}^{y_0} f(y)dy$$

$$P(y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} f(y)dy = F(y_2) - F(y_1)$$

$$E(g(y)) = \int_{-\infty}^{\infty} g(y) f(y)dy$$

$$\mu = E(y) = \int_{-\infty}^{\infty} yf(y)dy$$

paradox? $P(Y = y_0) = 0$

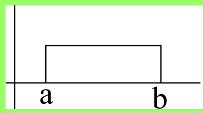
since

$$\int_{y_0}^{y_0} f(y)dy = 0$$

$$P(y_1 \leq Y \leq y_2) = P(y_1 < Y < y_2)$$

Uniform Random Variable

a special continuous random variable



$$f(y) = \begin{cases} \dfrac{1}{b-a} & a < Y < b \\ 0 & \text{otherwise} \end{cases}$$

If Y is Uniform then

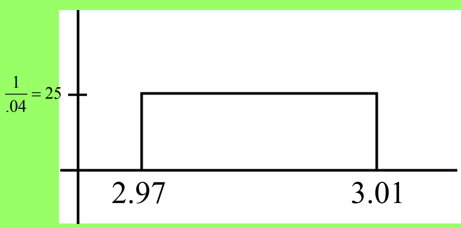
$$\mu = E(Y) = \frac{a+b}{2}$$

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

$$\sigma = SD(Y) = \frac{b-a}{\sqrt{12}}$$

Y = weight (grams) of a randomly selected car part

Y is Uniform on (2.97, 3.01)



On average the car parts weigh

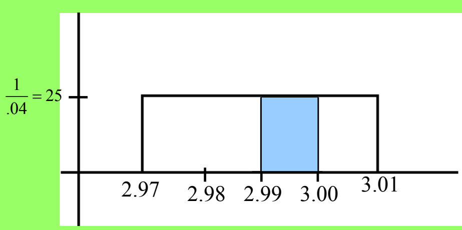
$$\mu = \frac{2.97 + 3.01}{2} = 2.99 \text{ grams}$$

$$\sigma^2 = Var(Y) = \frac{(3.01 - 2.97)^2}{12} = \frac{1}{7500}$$

$$\sigma = SD(Y) = \frac{1}{\sqrt{7500}}$$

$$P(2.99 < Y < 3.00) = .01(25) = .25$$

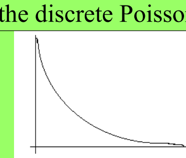
$$P(2.99 < Y < 3.00) = \int_{2.99}^{3.00} 25dy = .25$$



Exponential Random Variable

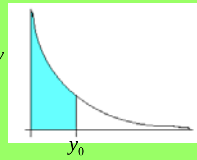
a special continuous random variable

(closely related to the discrete Poisson r.v.)



$$f(y) = \begin{cases} \lambda e^{-\lambda y} & y > 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(y) = \lambda e^{-\lambda y}$$



$$F(y_0) = P(Y \le y_0) = \int_{-\infty}^{y_0} f(y)dy = \int_{0}^{y_0} \lambda e^{-\lambda} dy =$$

$$\left(-e^{-\lambda(y_0)}\right) - \left(-e^{\lambda(0)}\right) = 1 - e^{-\lambda y_0}$$

If Y is Exponential then

$$\boxed{\mu = E(Y) = \int_0^{\infty} y\lambda e^{-\lambda y} dy = \frac{1}{\lambda}}$$

$$E(Y^2) = \int_0^{\infty} y^2 \lambda e^{-\lambda y} dy = \frac{2}{\lambda^2}$$

$$\boxed{\sigma^2 = Var(Y) = E(Y^2) - \mu^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}}$$

$$\boxed{\sigma = SD(Y) = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}}$$

<u>recall</u>: a Poisson r.v. counts the # of occurrences of an event in a specific time period

Y = { 0, 1, 2, … }

λ = mean rate of occurrences

<u>now</u>: an Exponential r.v. is the waiting time until the next occurrence

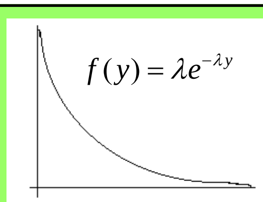
<u>Poisson</u>: # of lightning strikes to ME-EM per year

$$\lambda = \frac{1}{8}$$

<u>Exponential</u>: waiting time until next strike

On average, we wait 8 years until the next strike:

$$\mu = E(Y) = \frac{1}{\lambda} = \frac{1}{1/8} = 8$$



$$f(y) = \begin{cases} \frac{1}{8} e^{-\frac{1}{8}y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

<u>before</u>: (Poisson)

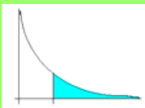
P(no lightning strikes for the next year) =

$$P(Y = 0) = \frac{\left(\frac{1}{8}\right)^0}{0!} e^{-\frac{1}{8}} = e^{-\frac{1}{8}} = .88250$$

Poisson with $\lambda = \frac{1}{8}$

$$\boxed{P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}}$$

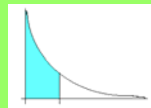
now: (Exponential)
P(waiting time is one year or more) =


1

$$P(Y>1)=\int_{1}^{\infty}\frac{1}{8}e^{-\frac{1}{8}y}\,dy=$$

$$=\left(-e^{-\frac{1}{8}(\infty)}\right)-\left(-e^{-\frac{1}{8}(1)}\right)=e^{-\frac{1}{8}}=.88250$$

P(we will have a lightning strike within the next 5.5 years) =


5.5

$$P(Y<5.5)=\int_{0}^{5.5}\frac{1}{8}e^{-\frac{1}{8}y}\,dy=$$

$$=\left(-e^{-\frac{1}{8}(5.5)}\right)-\left(-e^{-\frac{1}{8}(0)}\right)=1-e^{-\frac{1}{8}(5.5)}=.49717$$

Y = waiting time (minutes) until the next customer logs in to their account on a bank website

Y is Exponential with $\lambda = 2$

On average, we wait a half minute until the next customer logs on:

$$\mu = E(Y)=\frac{1}{\lambda}=\frac{1}{2}$$

(Poisson would count how many customer log-on's per minute)

Y is Exponential with $\lambda = 2$

$$f(y)=\begin{cases} 2e^{-2y} & y>0 \\ 0 & \text{otherwise} \end{cases}$$

P(no log-on's for the next 5 minutes) =


5

P(no log-on's for the next 5 minutes) =


5

$$P(Y>5)=\int_{5}^{\infty}2e^{-2y}\,dy=$$

$$=\left(-e^{-2(\infty)}\right)-\left(-e^{-2(5)}\right)=e^{-10}=.0000454\approx\frac{1}{22026}$$

Two other continuous r.v.'s found in the text:

Weibull
Gamma

MA 3710

3.5 – 3.7

Eiffel Tower
Paris

Paris Casino
Las Vegas

	score	possible	μ	σ	Z
You	100	200	80	20	1
Friend	75	100	80	10	–0.5

$$Z = \frac{100-80}{20} = 1$$

$$Z = \frac{75-80}{10} = -0.5$$

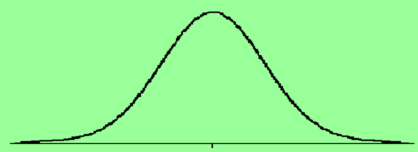
$$Z = \frac{Y-\mu}{\sigma}$$

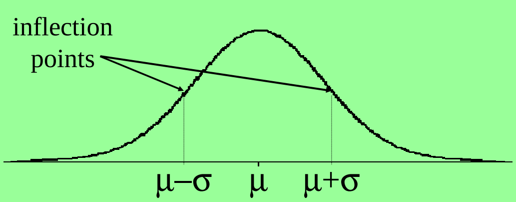
"standardizing" an observation (finding its Z-score) is measuring how many standard deviations above or below the mean it is

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

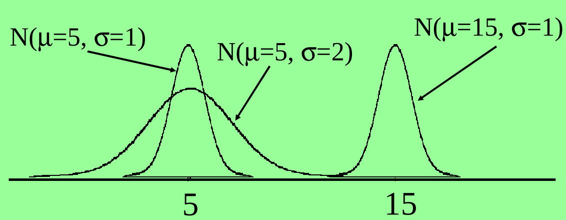
Normal distribution (bell curve)
The most famous continuous r.v.

If Y is normally distributed, its possible values are $(-\infty, \infty)$. The curve is centered around its mean μ and its width depends on its standard deviation σ



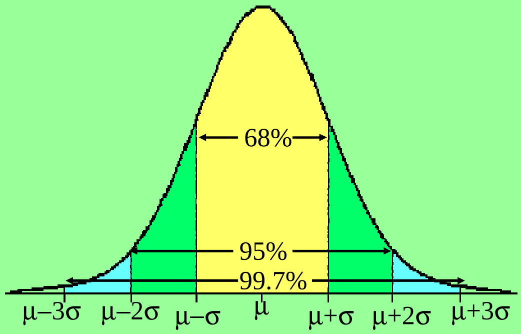


$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \qquad -\infty < y < \infty$$



Would a person with a height 10σ above the mean be "pretty tall", or would they be a giant?

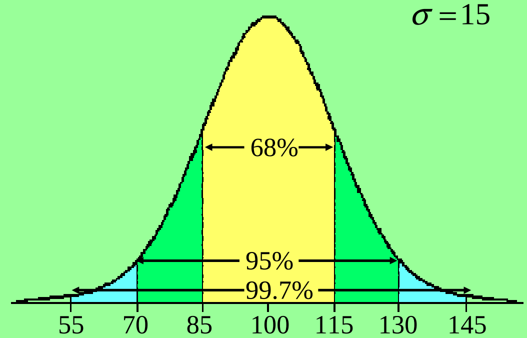
Empirical Rule for normal distributions



IQ scores

$\mu = 100$

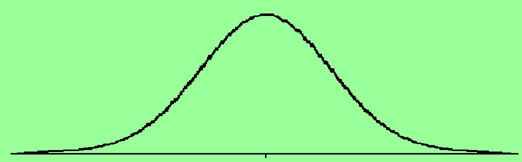
$\sigma = 15$



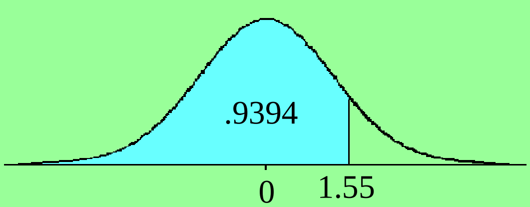
if a normal curve has μ=0 and σ=1, it's called the standard normal curve and we use Z instead of Y.

Z ~ N(μ=0, σ=1)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty$$



P(Z < 1.55)

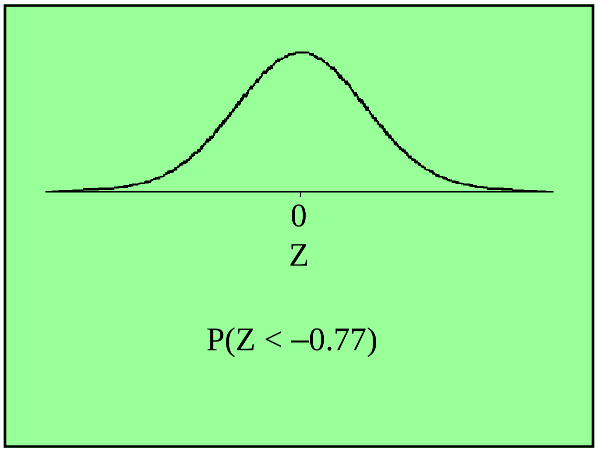


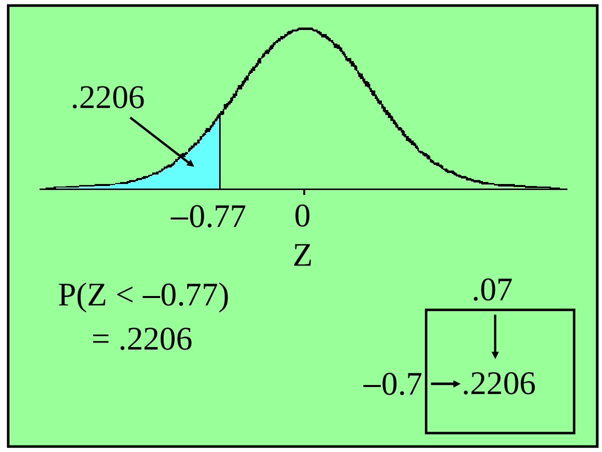
P(Z < 1.55)
= .9394

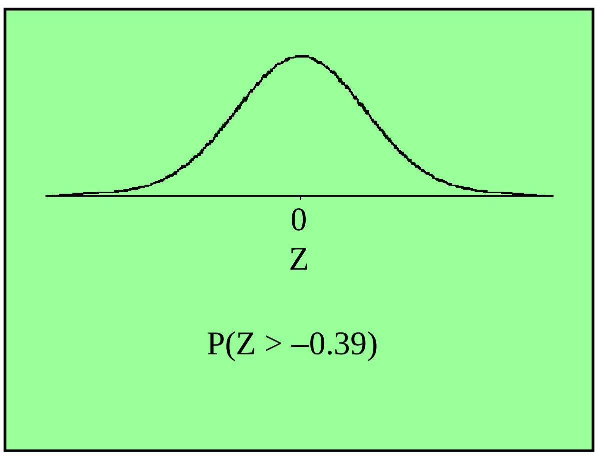
	.05
1.5	→ .9394

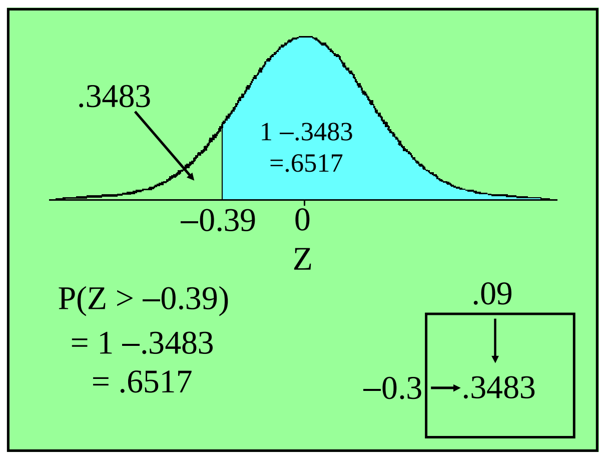
Standard Normal Table

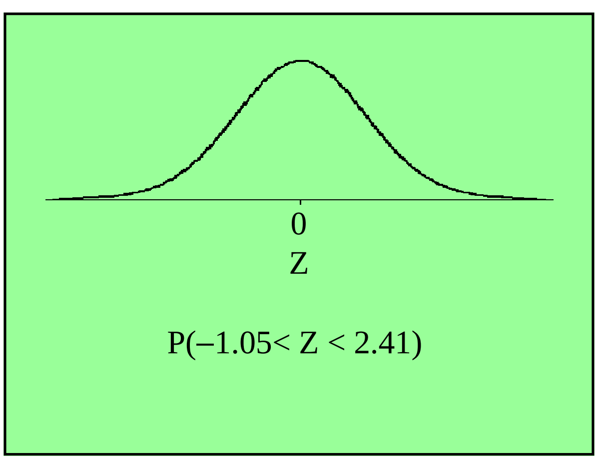
1. area under curve = 1
2. curve is symmetric around zero
3. table always gives you area to the LEFT of a z value

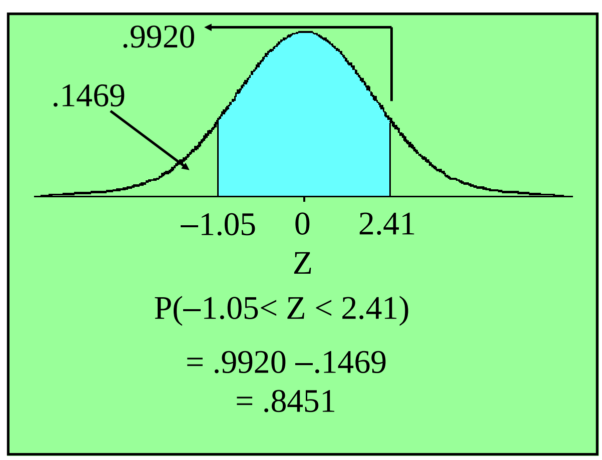

0
Z
P(Z < –0.77)

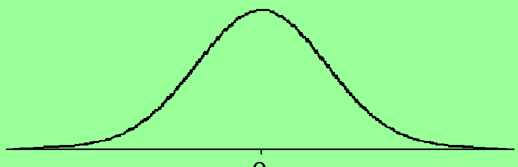

.2206
–0.77
0
Z
P(Z < –0.77)
= .2206
.07
–0.7
.2206


0
Z
P(Z > –0.39)

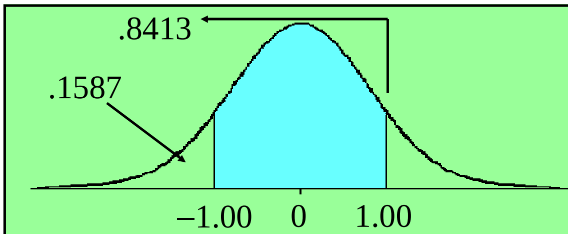

.3483
1 –.3483
=.6517
–0.39
0
Z
P(Z > –0.39)
= 1 –.3483
= .6517
.09
–0.3
.3483


0
Z
P(–1.05< Z < 2.41)


.9920
.1469
–1.05
0
2.41
Z
P(–1.05< Z < 2.41)
= .9920 –.1469
= .8451



P(–1.00< Z < 1.00)

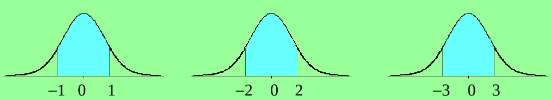


P(–1.00< Z < 1.00)

= .8413 –.1587

= .6826

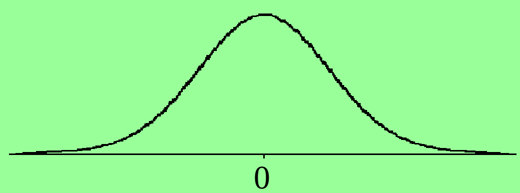
Empirical Rule!



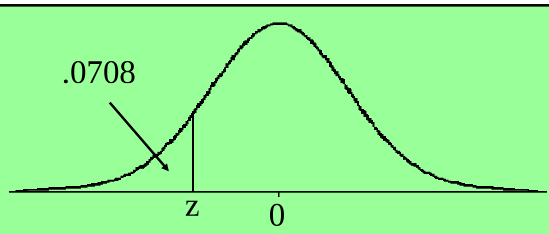
P(–1< Z < 1) = .6826 ≈ 68%

P(–2< Z < 2) = .9544 ≈ 95%

P(–3< Z < 3) = .9974 ≈ 99.7%

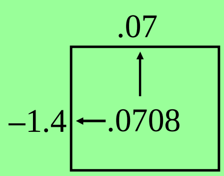


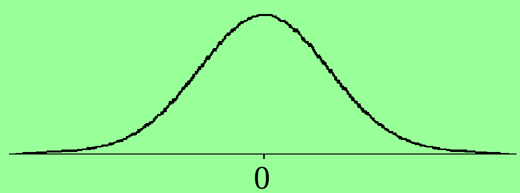
P(Z < z) = .0708



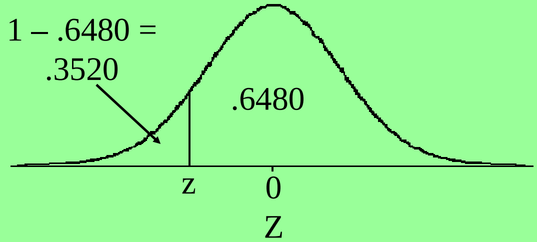
P(Z < z) = .0708

z = –1.47





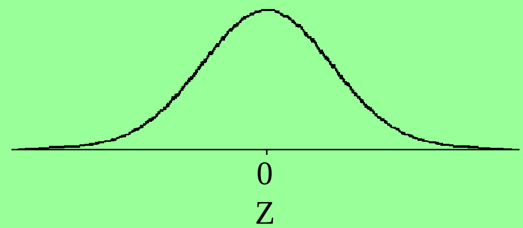
P(Z > z) = .6480



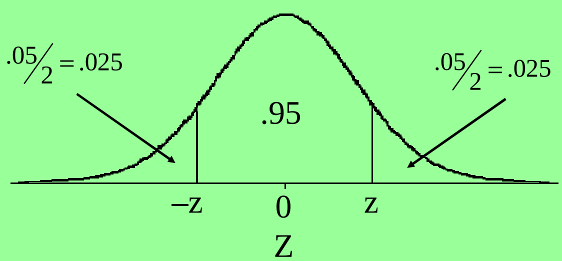
P(Z > z) = .6480

z = –0.38

	.08
–0.3	.3520



P(–z < Z < z) = .95

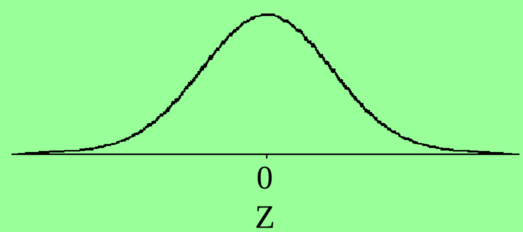


P(–z < Z < z) = .95

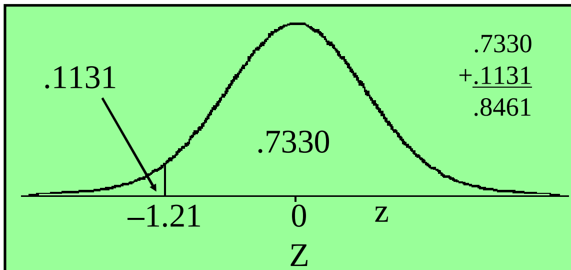
z = 1.96

–z = –1.96

	.06
–1.9	.0250



P(–1.21 < Z < z) = .7330



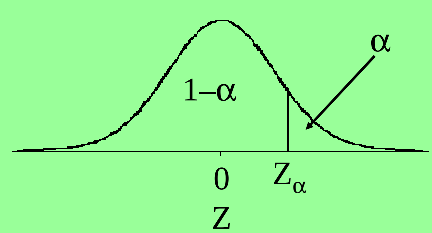
.7330
+.1131
.8461

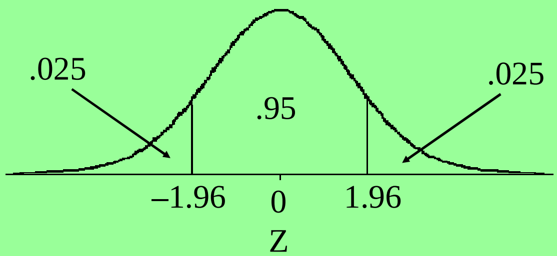
P(–1.21 < Z < z) = .7330

z = 1.02

	.02
1.0	.8461

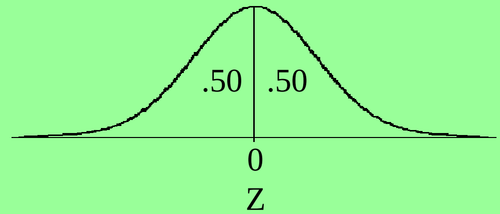
Z_α = value with α area to its right





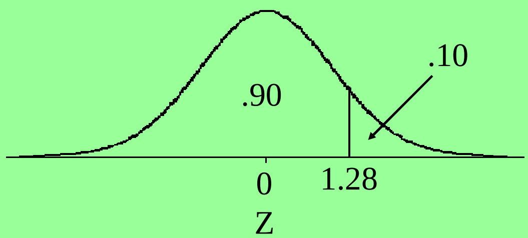
$Z_{.025} = 1.96$ $\quad$ $Z_{.975} = -1.96$

$Z_{.50} = 0$



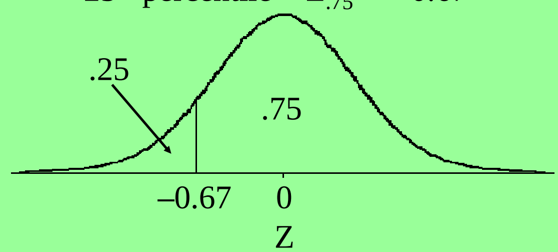
$Z_{\alpha} = (1-\alpha)100^{th}$ percentile

90^{th} percentile = $Z_{.10} = 1.28$



10^{th} percentile = $Z_{.90} = -1.28$

25^{th} percentile = $Z_{.75} = -0.67$



75^{th} percentile = $Z_{.25} = 0.67$

"regular" normal distributions

standard normal = $Z \sim N(\mu=0, \sigma=1)$

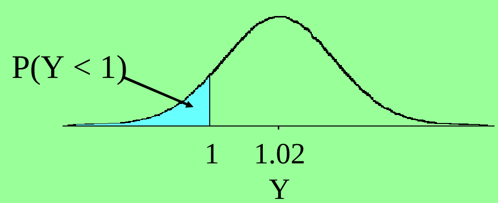
(we have table)

normal = $Y \sim N(\mu, \sigma)$

(no table!)

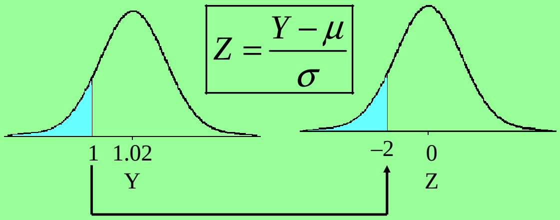
Y = gallons of milk placed in a plastic jug

$Y \sim N(\mu=1.02, \sigma=.01)$



$$\boxed{Z = \frac{1-1.02}{.01} = -2}$$

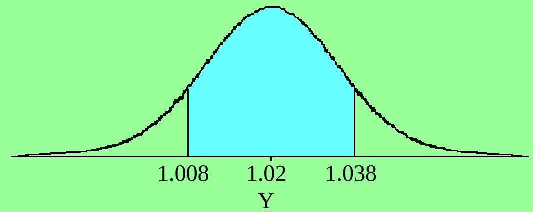
$$Z = \frac{Y-\mu}{\sigma}$$



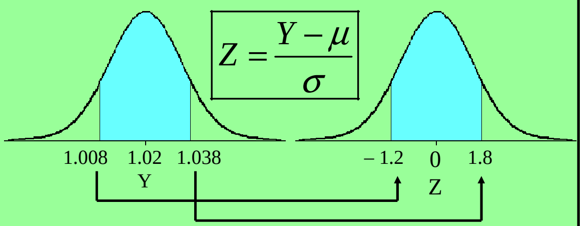
P(Y < 1) = $P\left(Z < \frac{1-1.02}{.01}\right) =$

P(Z < –2) = .0228

P(1.008 < Y < 1.038) =



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

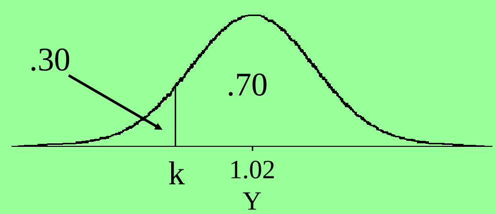


P(1.008<Y<1.038) =

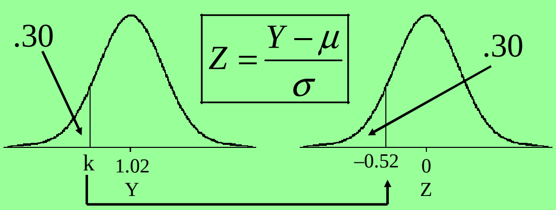
$$P\left(\frac{1.008-1.02}{.01} < Z < \frac{1.038-1.02}{.01}\right)$$

= P(–1.20< Z < 1.80) = .9641 –.1151 = .8490

70% of all milk jugs contain more than k gallons. (k is the 30th percentile). what is k?



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



$$Z = \frac{k-1.02}{.01} = -0.52$$

k = 1.0148

$$\frac{k-1.02}{.01} = -0.52$$

$$k - 1.02 = -0.52(.01)$$

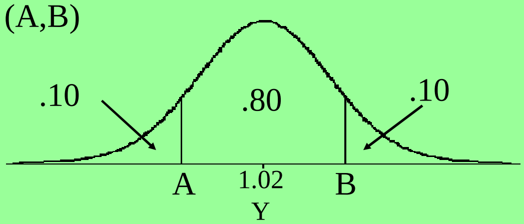
$$k = 1.02 - 0.52(.01) = 1.0148$$

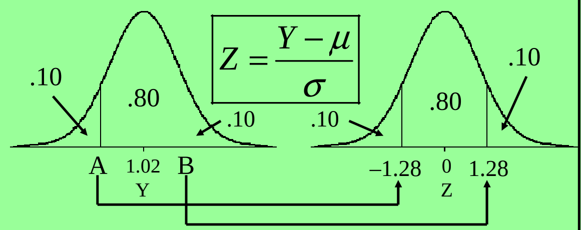
percentile on any normal curve = μ + Zσ

corresponding percentile on Z-curve

80% of all milk jugs contain a volume of milk in an interval centered around μ=1.02. What is it?

(A,B)





$$\frac{A-1.02}{.01} = -1.28 \qquad \frac{B-1.02}{.01} = 1.28$$

A = 1.0072 B = 1.0328

(1.0072, 1.0328)

$\overline{Y}$ is a random variable

$$\mu_{\overline{Y}} = E\left(\overline{Y}\right) = \mu$$

(the sample mean is an unbiased estimator for the population mean)

$$\sigma^2_{\overline{Y}} = Var\left(\overline{Y}\right) = \frac{\sigma^2}{n}$$

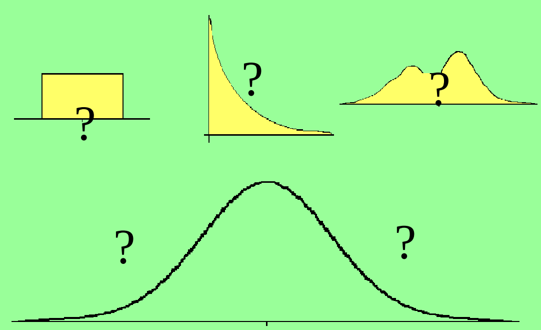
$$\sigma_{\overline{Y}} = SD\left(\overline{Y}\right) = \frac{\sigma}{\sqrt{n}}$$

IF sampling with replacement or from an infinitely large population

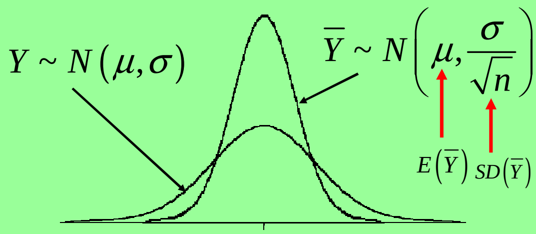
We know the mean, variance, and standard deviation of the sample mean.

But what kind of random variable is it???

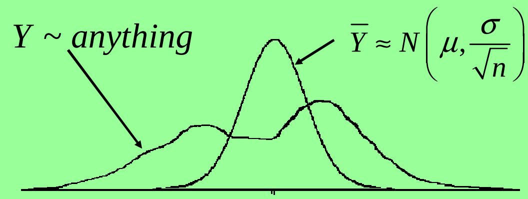
Central Limit Theorem



Case 1: if you are sampling from a normal population, then the sample mean is normally distributed



Case 2: even if the population isn't normal, the sample mean is still approximately normally distributed if the sample size is large (n ≥ 30)



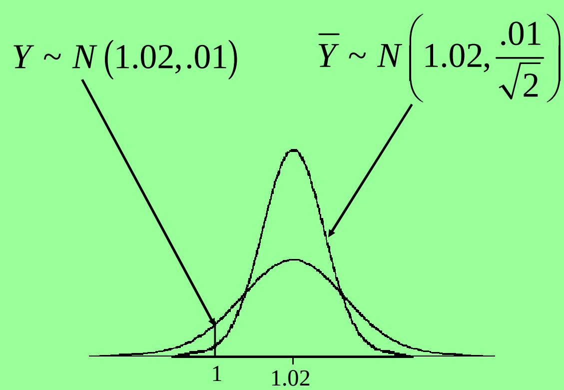
Y = gallons of milk placed in a plastic jug

Y ~ N(μ=1.02, σ=.01)

before: n = 1 : P(Y < 1) = .0228

now: What if we randomly select n = 2 jugs and compute the sample mean?

$$P(\bar{Y} < 1) = ?$$

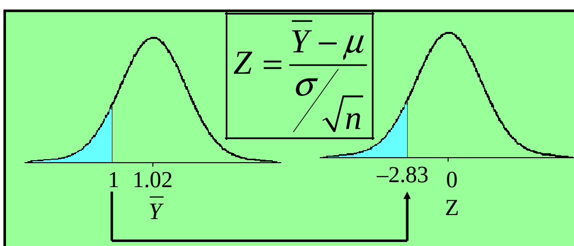


$$Z = \frac{Y - \mu}{\sigma}$$

before, when we were interested in Y

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

now we are interested in $\bar{Y}$

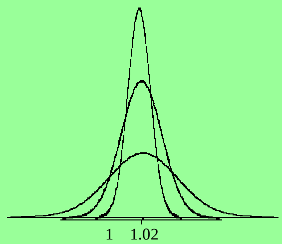


$$P(\bar{Y} < 1) = P\left(Z < \frac{1 - 1.02}{.01/\sqrt{2}}\right) =$$

P(Z < –2.83) = .0023

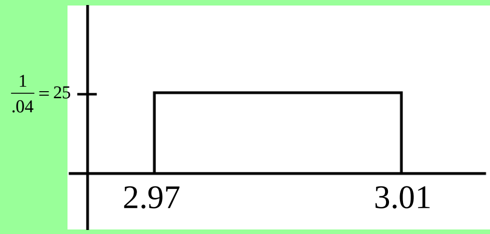
As n increases, the probability decreases:

n	Z	$P(\bar{Y}<1)$
1	–2	.0228
2	–2.83	.0023
3	–3.46	$\approx \frac{1}{3759}$
4	–4	$\approx \frac{1}{31574}$
5	–4.47	$\approx \frac{1}{258257}$



<u>from before</u>: Y = weight (grams) of a randomly selected car part

Y is Uniform on (2.97, 3.01)



On average the car parts weigh

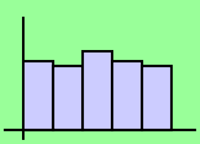
$$\mu = \frac{2.97+3.01}{2} = 2.99 \text{ grams}$$

$$\sigma^2 = Var(Y) = \frac{(3.01-2.97)^2}{12} = \frac{1}{7500}$$

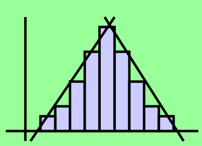
$$\sigma = SD(Y) = \frac{1}{\sqrt{7500}}$$

If you select n = 30 car parts at random and compute the sample mean, find

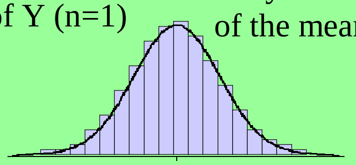
$$P(2.987 < \bar{Y} < 2.991)$$



many observations of Y (n=1)

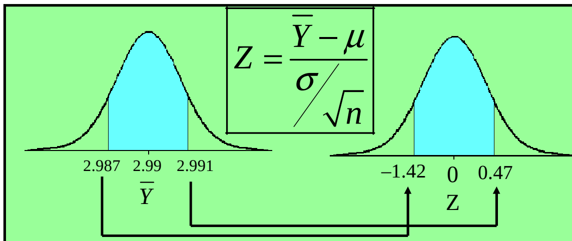


many observations of the mean (n=5)



many observations of the mean (n=30)

$$Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$$



$$P(2.987<\bar{Y}<2.991) = P\left(\frac{2.987-2.99}{\left(\frac{1}{\sqrt{7500}}\right)/\sqrt{30}} < Z < \frac{2.991-2.99}{\left(\frac{1}{\sqrt{7500}}\right)/\sqrt{30}}\right) =$$

$$P(2.987 < \overline{Y} < 2.991) =$$

$$P\left(\frac{2.987-2.99}{\left(\frac{1}{\sqrt{7500}}\right)\Big/\sqrt{30}} < Z < \frac{2.991-2.99}{\left(\frac{1}{\sqrt{7500}}\right)\Big/\sqrt{30}}\right) =$$

P(–1.42 < Z < 0.47) =

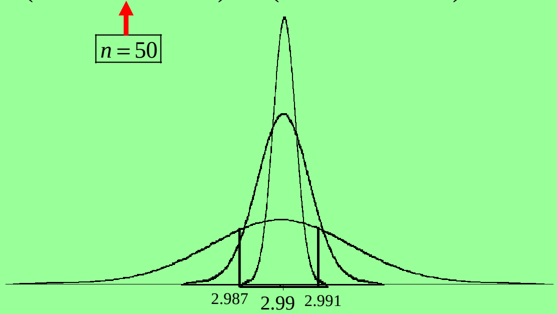
.6808 – .0778 = .6030

$$P(2.987 < \underset{\boxed{n=30}}{\overline{Y}} < 2.991) = .6030$$

On your own later:

$$P(2.987 < \underset{\boxed{n=50}}{\overline{Y}} < 2.991) = ?$$

$$P(2.987 < \underset{\boxed{n=50}}{\overline{Y}} < 2.991) = P(-1.84 < Z < 0.61) = .6932$$



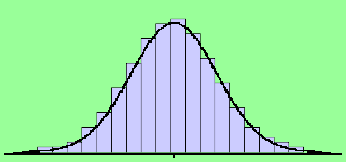
Assessing Normality

How can I tell if my data came from a normal distribution?

If your data is normal, then…

---

1. Histogram and stem-and-leaf plots should be bell-shaped:



```
0.1 | 0
0.2 | 7
0.3 | 33
0.4 | 0125
0.5 | 456678
0.6 | 14667899
0.7 | 224456
0.8 | 1132
0.9 | 69
1.0 | 2
1.1 | 0
```

---

2. Empirical Rule:

About 68% of the data should be in:

$$(\overline{Y} - s, \overline{Y} + s)$$

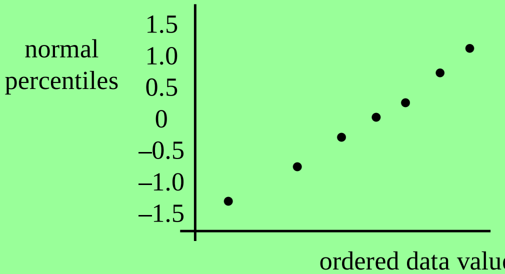
About 95% of the data should be in:

$$(\overline{Y} - 2s, \overline{Y} + 2s)$$

About 99.7% of the data should be in:

$$(\overline{Y} - 3s, \overline{Y} + 3s)$$

3. Normal Probability Plot (QQ Plot) should look linear:

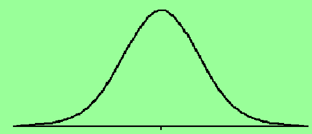


suppose n = 10

10 randomly selected football players weights (ordered)

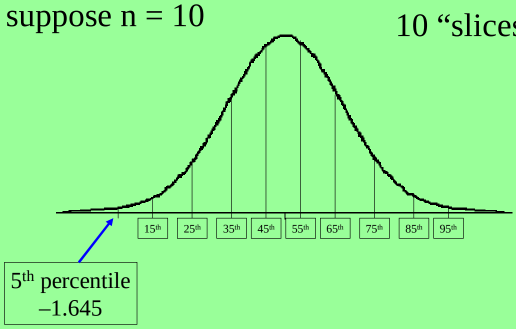
{ 248, 253, 263, 272, 274, 286, 307, 310, 312, 323}

Is the population of weights normal?

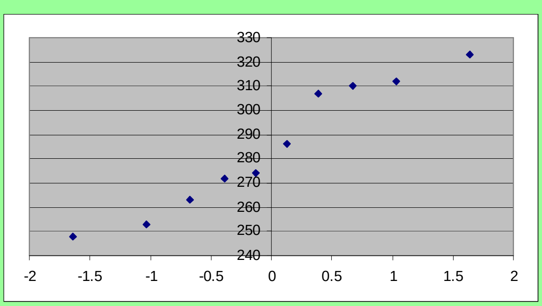


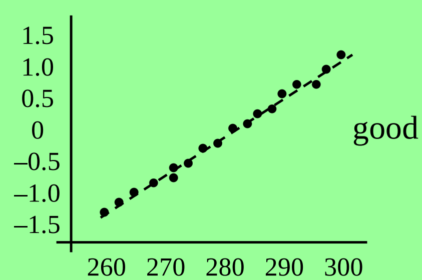
suppose n = 10

10 “slices”

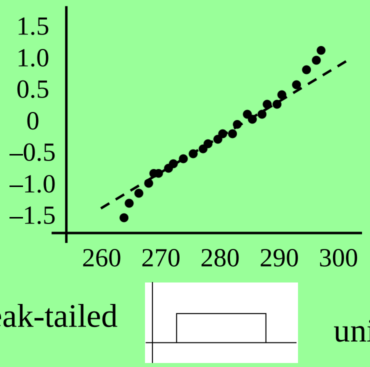


percentile	Z	weight
5th	–1.645	248
15th	–1.04	253
25th	–0.67	263
35th	–0.39	272
45th	–0.13	274
55th	0.13	286
65th	0.39	307
75th	0.67	310
85th	1.04	312
95th	1.645	323

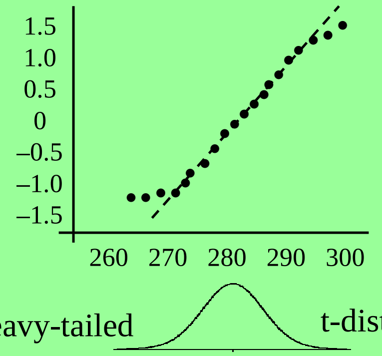




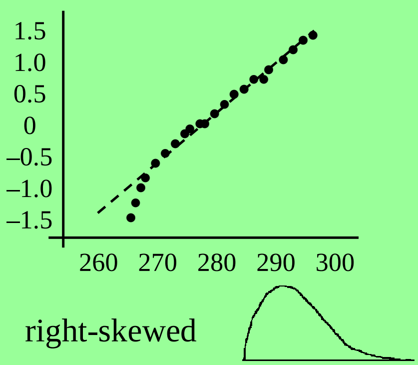
if the plot is roughly linear, it is reasonable to assume the variable is at least approximately normal



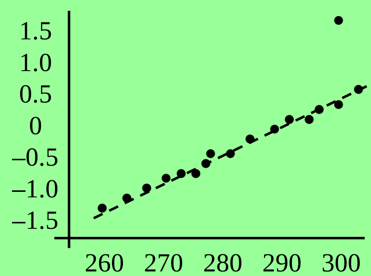
weak-tailed uniform



heavy-tailed t-distribution



right-skewed



watch for outliers

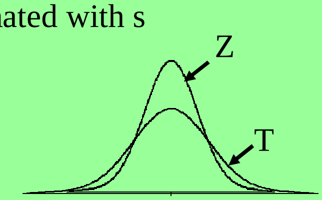
If $Y \sim N(\mu, \sigma)$ then

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

one random variable

If σ is unknown, then it must be estimated with s

$$T = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$



two random variables

data set: $\{Y_1, Y_2, \ldots, Y_n\}$

The sample variance examines deviations (distances from values to the mean)

$$\left(Y_1-\overline{Y}\right),\left(Y_2-\overline{Y}\right),\ldots,\left(Y_n-\overline{Y}\right)$$

$$\left(Y_1-\overline{Y}\right)+\left(Y_2-\overline{Y}\right)+\ldots+\left(Y_n-\overline{Y}\right)=$$

$$\sum Y - n\overline{Y} = \sum Y - \cancel{n}\frac{\sum Y}{\cancel{n}} = 0$$

The <u>sample variance</u> is the "average" squared deviation

sample variance = s^2

$$s^2 = \frac{\sum (Y-\overline{Y})^2}{n-1} = \frac{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}{n-1} = \frac{n\sum Y^2 - \left(\sum Y\right)^2}{n(n-1)}$$

regular convenient book

sample standard deviation = s = $\sqrt{s^2}$

data set: { 3, 7, 10, 11, 14 }

$$\overline{Y} = \frac{45}{5} = 9$$

Y	$Y-\overline{Y}$	$\left(Y-\overline{Y}\right)^2$	Y^2
3	–6	36	9
7	–2	4	49
10	1	1	100
11	2	4	121
14	5	25	196
45	0	70	475

$$s^2 = \frac{\sum (Y-\overline{Y})^2}{n-1} = \frac{70}{4} = 17.5$$

$$\boxed{s = \sqrt{17.5} = 4.18}$$

$$s^2 = \frac{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}{n-1} = \frac{475 - \frac{45^2}{5}}{4} = 17.5$$

$$s^2 = \frac{n\sum Y^2 - \left(\sum Y\right)^2}{n(n-1)} = \frac{5(475)-45^2}{5(4)} = 17.5$$

A <u>standard deviation</u> is the "typical distance" from a value to the mean in a data set

{ 10, 10, 10, 12, 12, 12 }

$$\overline{Y} = 11, s \approx 1, s = 1.095$$

{ 0, 0, 0, 0, 100, 100, 100, 100 }

$$\overline{Y} = 50, s \approx 50, s = 53.45$$

{ 3, 7, 10, 11, 14 } $\overline{Y} = 9, s^2 = 17.5, s = 4.18$

{ 13, 17, 20, 21, 24 } $\overline{Y} = 19, s^2 = 17.5, s = 4.18$

If you add k to a data set:

$$\overline{Y}_{new} = \overline{Y}_{old} + k$$

$$s^2_{new} = s^2_{old}$$

$$s_{new} = s_{old}$$

{ 3, 7, 10, 11, 14 } $\overline{Y} = 9, s^2 = 17.5, s = 4.18$

{ 30, 70, 100, 110, 140 } $\overline{Y} = 90$

$s^2 = (10)^2 \times 17.5 = 1750$

$s = (10)4.18 = 41.8$

If you multiply a data set by k:

$\overline{Y}_{new} = \left(\overline{Y}_{old}\right)k$

$s^2{}_{new} = \left(s^2{}_{old}\right)k^2$

$s_{new} = \left(s_{old}\right)|k|$

{ 3, 7, 10, 11, 14 } $\overline{Y} = 9, s^2 = 17.5, s = 4.18$

{ –3, – 7, – 10, – 11, – 14 }

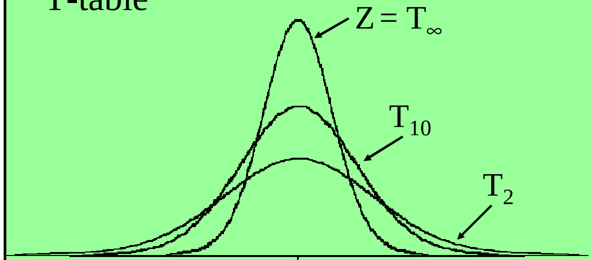
$\overline{Y}_{new} = \left(\overline{Y}_{old}\right)k$ $\quad \overline{Y} = 9(-1) = -9$

$s^2{}_{new} = \left(s^2{}_{old}\right)k^2$ $\quad s^2 = 17.5 \times (-1)^2 = 17.5$

$s_{new} = \left(s_{old}\right)|k|$ $\quad s = 4.18 \times |-1| = 4.18$

T-curve and T-table

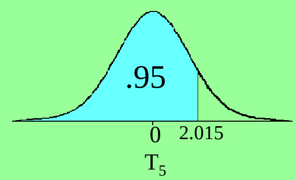
*symmetric around zero
*total area under curve = 1
*shorter and wider than Z
*degrees of freedom

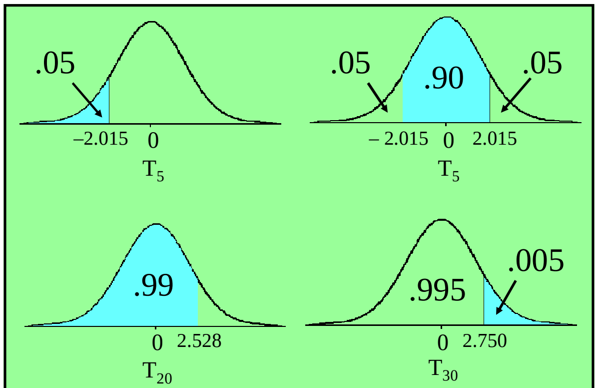


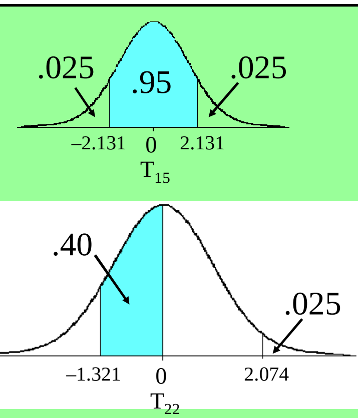
The T-table gives you t-values with an area to the left

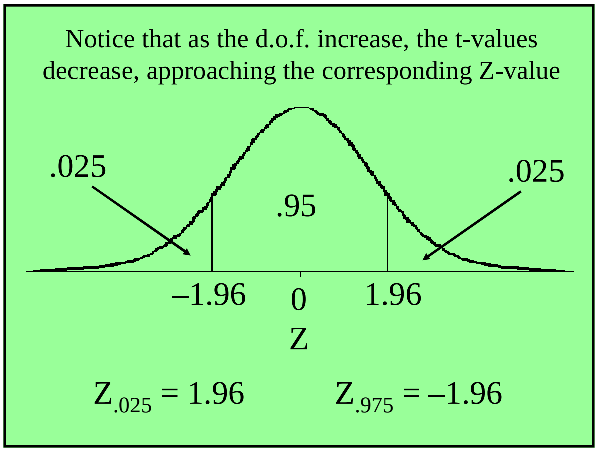
95%

5 → 2.015



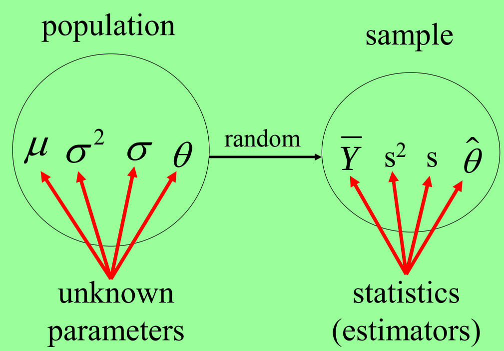





Notice that as the d.o.f. increase, the t-values decrease, approaching the corresponding Z-value
.025
.025
.95
–1.96
0
1.96
Z
Z.025 = 1.96
Z.975 = –1.96

MA 3710
CH 4


Parthenon
Athens



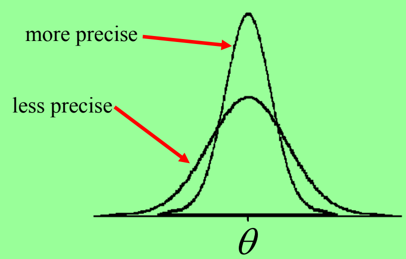
Statistics (estimators) are random variables with their own means, variances, and standard deviations.

If $E\left(\hat{\theta}\right)=\theta$ then $\hat{\theta}$ is an unbiased estimator for θ

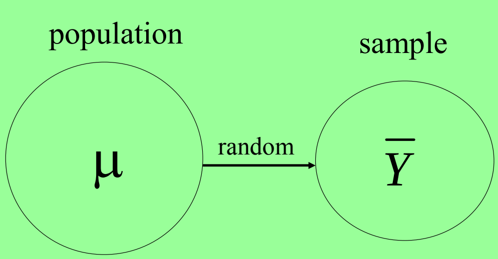
$\overline{Y}$ and s^2 are unbiased: $\boxed{\begin{array}{l} E\left(\overline{Y}\right)=\mu \\ E\left(s^2\right)=\sigma^2 \end{array}}$

s is biased: $\boxed{E(s)\neq\sigma}$

An estimator with a smaller variance is called more <u>precise</u>:



Inferential Statistics



<u>Three Options</u>:

1. Estimate μ with $\overline{Y}$ (point estimate)

2. calculate an interval of plausible values for μ

$\left(\overline{Y}-E,\overline{Y}+E\right)$

E = error margin

(confidence interval)

3. answer a question about μ

<u>example</u>: does $\mu = 8$? Yes/No

(hypothesis test)

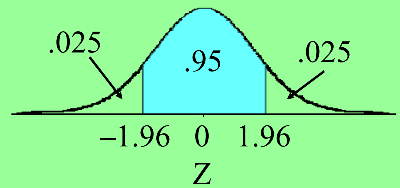
situation:

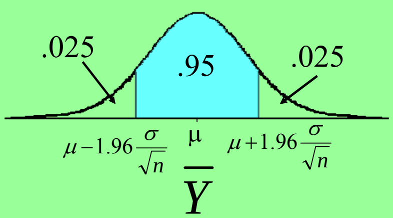
$\bar{Y}$ is normal because Y is normal, OR
$\bar{Y}$ is approx. normal because n≥30 (CLT)

σ is known (for now)

$$\bar{Y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$E\left(\bar{Y}\right)$ $\qquad$ $SD\left(\bar{Y}\right)$





$$.95 = P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{Y} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{Y} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{Y} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(-\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(-\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

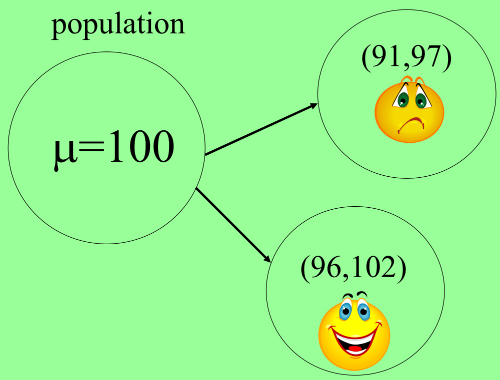
$$.95 = P\left(\bar{Y} + 1.96\frac{\sigma}{\sqrt{n}} > \mu > \bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$\boxed{.95 = P\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right)}$$

Therefore, if we take a random sample and calculate the interval

$$\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = \bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

then we are 95% confident μ is inside the interval
(the interval "captured" μ)



Interested in μ = population mean weight among all MTU students

$\bar{Y}$=150
n=30
known: σ=25

95% CI for μ

$$\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

$$150 \pm 1.96\frac{25}{\sqrt{30}} = 150 \pm 8.95$$

$$= \boxed{(141.05, 158.95)}$$

$$(141.05, 158.95)$$

True or false??

~~We sampled 95% of all MTU students.~~

false!

~~We know that μ=150.~~

false!

$$(141.05, 158.95)$$

True or false??

~~95% of all MTU students have a weight between 141.05 and 158.95~~

false!

~~We know that 141.05 < μ < 158.95~~

false!

$$(141.05, 158.95)$$

True or false??

~~P(141.05 < μ < 158.95) = .95~~

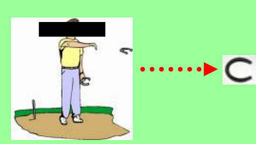
false!

we are 95% confident that
141.05 < μ < 158.95
(the interval captured μ)

true!

Analogy 1:
horseshoes

We know that in the long run, 95% of our CI's (horseshoes) are successful, but we aren't certain about any particular one.



Thrower can't see pin	*horseshoe*	*pin*
sampler doesn't know μ	CI	μ

Analogy 2: special coin

Building a CI is like tossing this coin off a bridge.

95% interval captured μ

5% interval didn't capture μ

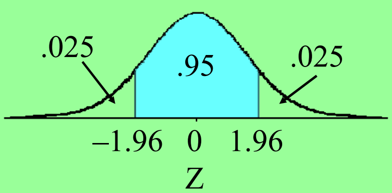
95% CI = $\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

confidence level

point estimate

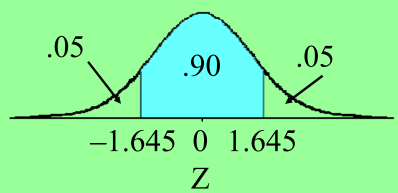
error margin

what about other confidence levels?...



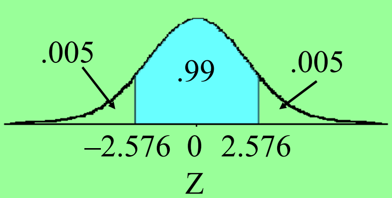
95% CI for μ =

$$\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$



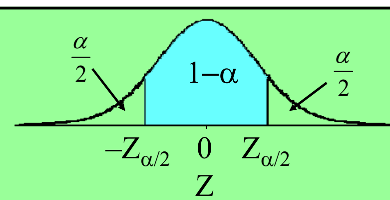
90% CI for μ =

$$\bar{Y} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$



99% CI for μ =

$$\bar{Y} \pm 2.576 \frac{\sigma}{\sqrt{n}}$$



If σ is known and $\bar{Y}$ is normal or approx. normal:
$(1-\alpha)100$ % CI for μ =

$$\boxed{\bar{Y} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

Thinner intervals are more useful than wider ones.
What makes an interval wider or thinner?

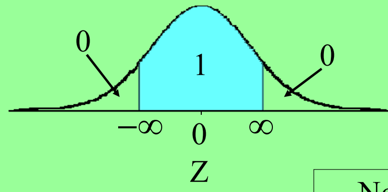
larger σ, wider interval

$$\bar{Y} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

more confidence, wider interval (trade off)

larger n, thinner interval

Would a 100% CI make sense?

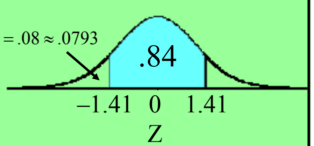


100% CI for μ =

$$\bar{Y} \pm \infty\frac{\sigma}{\sqrt{n}} = (-\infty, \infty)$$

No! It would be too wide to be useful.

$\bar{Y}$=50
n=100
known: σ=15



84% CI for μ $\quad \bar{Y} \pm 1.41\frac{\sigma}{\sqrt{n}}$

$$50 \pm 1.41\frac{15}{\sqrt{100}} = 50 \pm 2.115$$

$$= \boxed{(47.885, 52.115)}$$

Suppose (98,102) was a 95% CI for μ

what was $\bar{Y}$? $\quad \frac{98+102}{2} = 100$

what was the error margin?

$$100 \pm E = (98,102)$$

$$E = 2$$

---

interpretation: 95% of the time, the difference between $\bar{Y}$ and μ (the sampling error) is at most 2

if we want the EM to equal some small number B, what sample size is necessary?

$$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = B \longrightarrow \boxed{n = \left(\frac{Z_{\alpha/2}\sigma}{B}\right)^2}$$

<u>from before:</u> weights of MTU students

$\bar{Y}$=150
n=30
σ=25

95% CI for μ

$$= 150 \pm 8.95$$

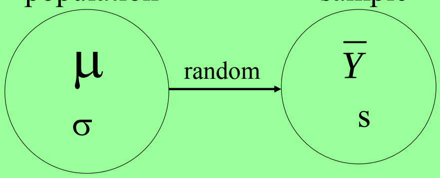
If we want an error margin of 3 pounds or less with 95% confidence, we should sample

$$n = \left(\frac{Z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.96(25)}{3}\right)^2 = 266.78 \uparrow 267$$

267 students

Up until now, we have assumed σ is known. More realistic: what if σ is unknown?

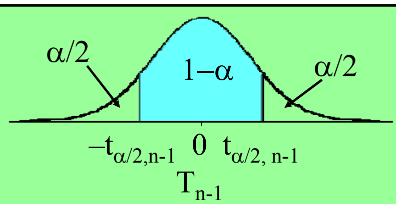
population | sample



$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

σ known → use Z-curve

σ unknown → use T-curve



If σ is unknown and $\bar{Y}$ is normal or approx. normal:
(1−α)100 % CI for μ =

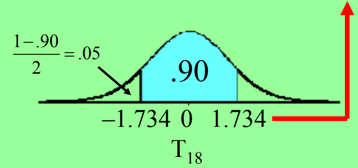
$$\boxed{\bar{Y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}}$$

μ = population mean cholesterol level among 70-year-old American men

$\bar{Y}$=44.2
n=19
s=3.8

90% CI for μ

$$\bar{Y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$



$\bar{Y}$=44.2
n=19
s=3.8

90% CI for μ

$$\bar{Y} \pm 1.734\frac{s}{\sqrt{n}}$$

$$44.2 \pm 1.734\left(\frac{3.8}{\sqrt{19}}\right) = 44.2 \pm 1.512$$

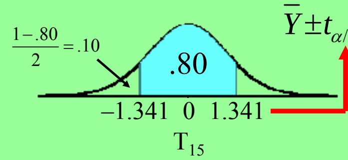
$$= \boxed{(42.688, 45.712)}$$

n = 16, σ unknown
An 80% CI for μ was (48,54)

what was $\bar{Y}$? $\quad \frac{48+54}{2} = 51$

what was s?

$$\bar{Y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$



$$51 \pm E = (48, 54) \rightarrow E = 3$$

$$\downarrow$$

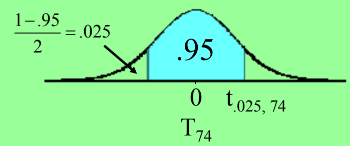
$$1.341 \frac{s}{\sqrt{n}} = 3$$

$$\downarrow$$

$$1.341 \frac{s}{\sqrt{16}} = 3 \rightarrow \boxed{s = 8.949}$$

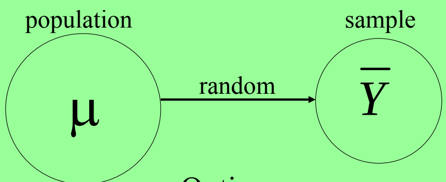
---

n = 75, σ unknown
want a 95% CI for μ



$t_{.025,\,74}$ is not on the table, so take $t_{.025,\,60} = 2.000$

If the d.o.f. you want is not on the table, take the largest d.o.f. less than the one you want. This will make CI's slightly wider than they should be.



Options:

1. point estimate
2. confidence interval
3. hypothesis test

answer a question about μ

example: does $\mu = 8$? Yes/No

---

null hypothesis (H_0)
statement to be tested

alternative hypothesis (H_a)
the alternative to H_0
(if H_0 isn't true, what do I suspect might be true instead?)

Analogy: court trial

H_0: defendant is innocent ← benefit of the doubt
H_a: defendant is guilty

decision
"reject H_0" or "don't reject H_0"

we reject H_0 only if we have convincing evidence (significant data) that H_0 is false

---

reality (unknown to us)

decision

	H_0 true	H_0 false
reject H_0	Type 1 error	
don't reject H_0		Type 2 error

Type 1 error (more serious):
rejecting H_0 when its true

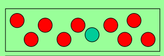
P(Type 1 error) = α = significance level

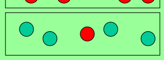
Type 2 error (less serious):
not rejecting H_0 when its false

P(Type 2 error) = β

<u>plan</u>: give H_0 benefit of the doubt (temporarily assume H_0 is true), examine the evidence (data) and see if it is consistent with H_0's claim

Three balls will be drawn at random with replacement from Basket A or Basket B. You don't know which one.

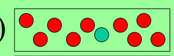
H_0: Basket A (9 red, 1 green) 

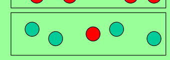
H_a: Basket B (1 red, 4 green) 

*let's reject H_0 if we observe <u>3 greens in a row</u>.

α = P(Type 1 error) = P(reject H_0 | H_0 is true) = P(GGG | Box A) =

$$(.1)^3 = .001 = \frac{1}{1000}$$

H_0: Basket A (9 red, 1 green) 

H_a: Basket B (1 red, 4 green) 

α = P(Type 1 error) = P(reject H_0 | H_0 is true) = P(GGG | Box A) = $(.1)^3 = .001 = \frac{1}{1000}$

β = P(Type 2 error) = P(don't reject H_0 | H_0 is false) = $P(\overline{GGG} \mid BoxB) = 1 - (.8)^3 = .488$

we could force $\alpha = 0$ by never rejecting H_0 but that would cause β to equal 1!

forcing α to be small usually causes β to be large and vice-versa (trade-off)

<u>example</u>

Purchasing shipments of iron ore pellets:
μ = mean weight among all shipments being purchased (unknown)
σ = 10 pounds (known)

Supplier claims that "the average shipment weighs 80 pounds"
You suspect you may be getting shortchanged.

H_0: μ = 80 pounds

three options for H_a

H_a: μ < 80 pounds (left-tailed or lower-tailed)

H_a: μ > 80 pounds (right-tailed or upper-tailed)

H_a: μ ≠ 80 pounds (two-tailed)

H_a: μ < 80 pounds

H_0: μ = 80

H_a: μ < 80

plan: randomly select n = 30 shipments, weigh them, compute $\overline{Y}$, and reject H_0 if $\overline{Y}$ is "too low".

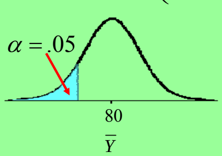
How low is "too low"?

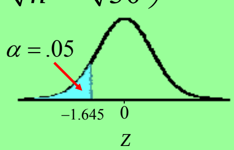
willing to make a Type 1 error with probability α = .05

---

*temporarily assume H_0 is true:

$$\overline{Y} \sim N\left(\mu = 80, \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{30}}\right)$$





$$Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}$$

---

*compute $\overline{Y}$, standardize (convert to a Z-score) and reject H_0 if Z < –1.645

rejection region/critical region:
R: Z < –1.645

*given info:
The experiment is conducted and $\overline{Y}$ = 76 pounds

---

test statistic: $Z = \dfrac{\overline{Y} - \mu}{\sigma / \sqrt{n}} = \dfrac{76 - 80}{10 / \sqrt{30}} = -2.19$

This test statistic is measuring how many standard deviations above or below average $\overline{Y}$ is if H_0 is true.

Since the test statistic is inside the rejection region, we reject H_0

What if, in reality, μ actually does equal 80 pounds?

then a Type 1 error occured

What if, in reality, μ equals 77.2 pounds?

then no error occured

One-sample Z-test for μ

$\overline{Y}$ is normal or approximately normal and σ is known

Step 1

choose H_0 , H_a , and α =significance level

H_0: μ = k

One-sample Z-test for μ

H_a: μ < k *OR* H_a: μ > k *OR* H_a: μ ≠ k

left-tailed lower-tailed | right-tailed upper-tailed | two tailed

---

Step 2

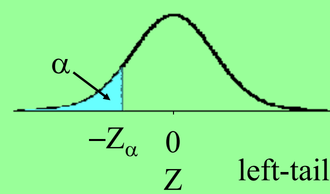
Determine test statistic:

$$Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}$$

Step 3

find R = rejection region/critical region

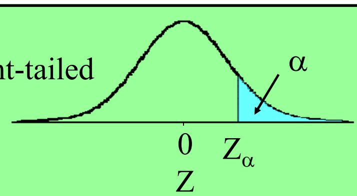
α → size; H_a → location

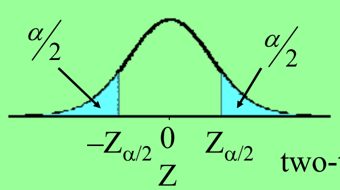


left-tailed

---

right-tailed





two-tailed

Step 4

Conduct experiment and find value of test statistic

Step 5

conclusion:

if the test statistic falls inside the rejection region, reject H_0

reject $H_0 \rightarrow$

"we have significant evidence at level α that H_0 is false"

don't reject $H_0 \rightarrow$

"data is not significant at level α"

Sometimes we compute a CI for μ after a rejection of H_0 to report the location of plausible values for μ.

example

H_0: $\mu = 50$

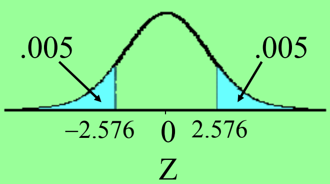
H_a: $\mu \neq 50$

$\alpha = .01$

$\sigma = 16$

$n = 64$

$\bar{Y} = 53$



R: $Z < -2.576$ or $Z > 2.576$

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} = \frac{53 - 50}{16 / \sqrt{64}} = 1.5$$

don't reject H_0

data is not significant at $\alpha = .01$

example

H_0: $\mu = 70$

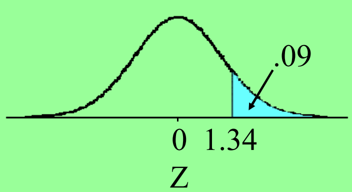
H_a: $\mu > 70$

$\alpha = .09$

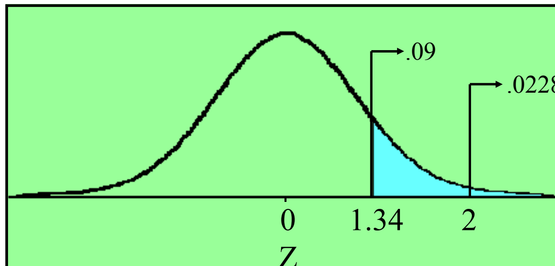
$\sigma = 12$

$n = 36$

$\bar{Y} = 74$

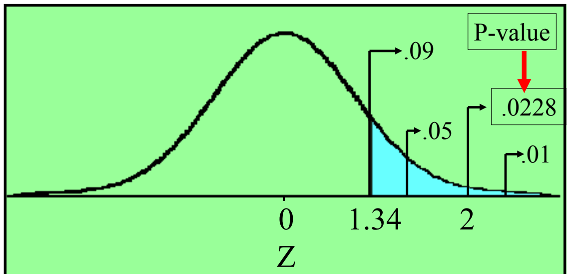


R: $Z > 1.34$



$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} = \frac{74 - 70}{12 / \sqrt{36}} = 2$$

reject H_0



Data significant at $\alpha = .09$
Data significant at $\alpha = .05$
Data not significant at $\alpha = .01$

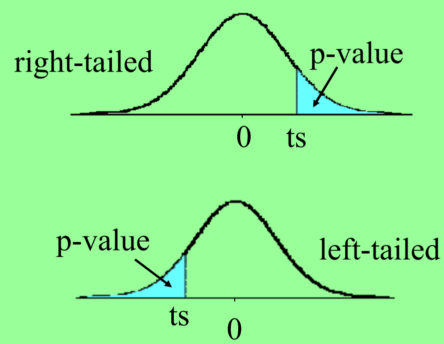
P-value

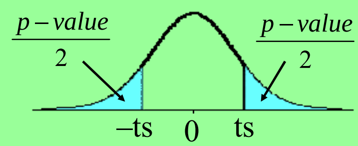
*p-value = area further into the tail(s) from the test stat.

*Measures how significant the data is

*The closer the p-value is to zero, the more evidence you have against H_0

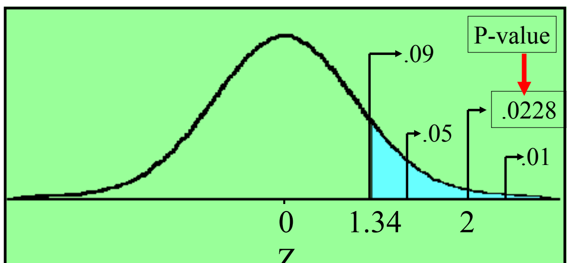
*If H_0 is true, p-value = the probability of observing data as extreme (or more extreme) as what was observed





two-tailed

On a two-tailed test, a p-value is twice the value it would otherwise be.



H_0 is rejected iff p-value < α

H_0 is not rejected iff p-value > α

$\alpha = .08$

Alice's p-value = .04
Betsy's p-value = $\frac{1}{1000000}$

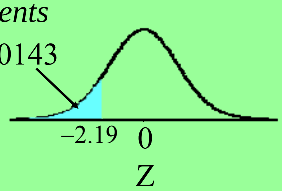
Both rejected H_0, but Betsy is more confident in her rejection than Alice is.

p-values in previous examples:

Iron ore shipments

H_0: $\mu = 80$
H_a: $\mu < 80$

$\alpha = .05$
$Z = -2.19$
rejected H_0

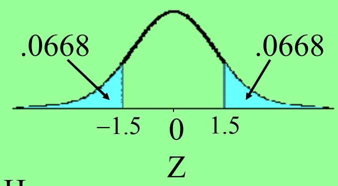


p-value = .0143 < α = .05

p-values in previous examples:

H_0: $\mu = 50$
H_a: $\mu \neq 50$

$\alpha = .01$
$Z = 1.5$
didn't reject H_0



p-value = 2(.0668) = .1336 > α = .01

How are hypothesis tests different when σ is unknown?

σ known → Z-curve
σ unknown → T-curve

A one-sample T-test for μ (σ unknown) is on a t-curve with n–1 d.o.f. and has a test statistic of

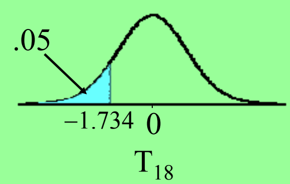
$$T = \frac{\overline{Y} - \mu}{s/\sqrt{n}}$$

μ = population mean cholesterol level among 70-year-old American men

H_0: $\mu = 45.5$
H_a: $\mu < 45.5$
$\alpha = .05$

$\overline{Y}$=44.2
n=19
s=3.8

<u>assume</u>: levels are normally distributed

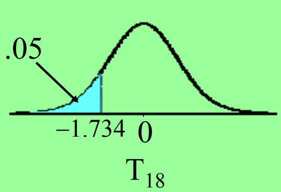


R: T < −1.734

$$T = \frac{\overline{Y} - \mu}{s/\sqrt{n}} = \frac{44.2 - 45.5}{3.8/\sqrt{19}} = -1.49$$

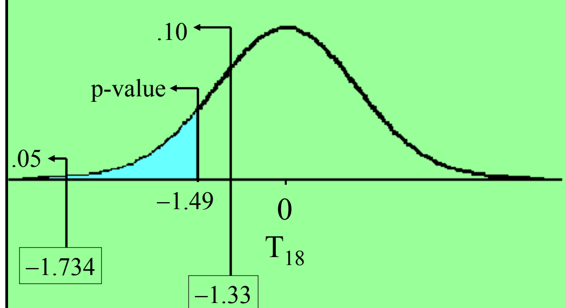
don't reject H_0

data not significant at $\alpha = .05$



R: T < −1.734

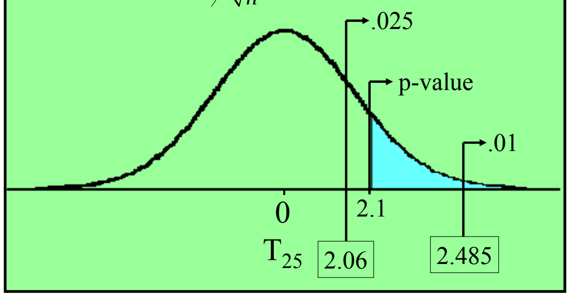
.05 < p-value < .10



Given these test statistics, place bounds on the p-value:

n=26
right-tailed

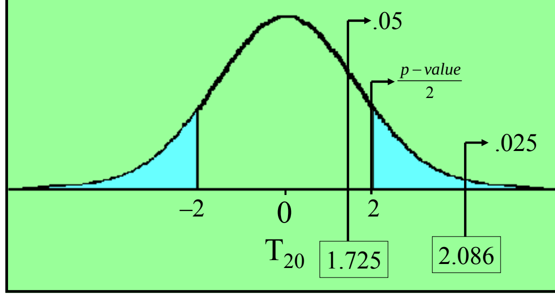
$T = \frac{\overline{Y} - \mu}{s/\sqrt{n}} = 2.1$ $\quad \boxed{.01 < \text{p-value} < .025}$



n=21
two-tailed

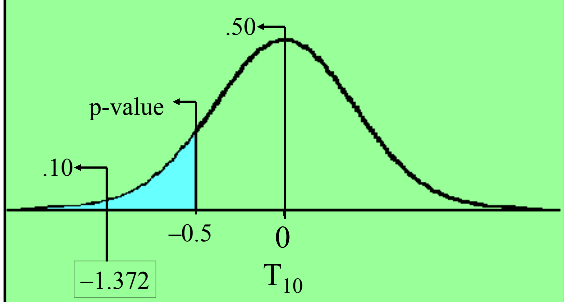
$T = \frac{\overline{Y} - \mu}{s/\sqrt{n}} = 2$ $\quad .025 < \frac{p - value}{2} < .05$

$\boxed{.05 < p - value < .10}$



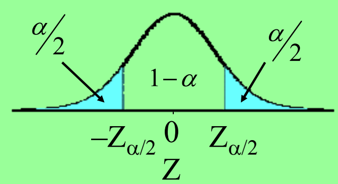
n=11
left-tailed

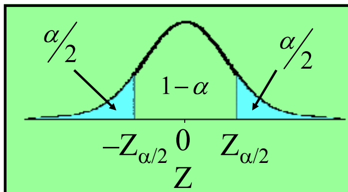
$T = \frac{\overline{Y} - \mu}{s/\sqrt{n}} = -0.5$ $\quad \boxed{.10 < p - value < .50}$



A connection between two-tailed tests and confidence intervals

H_0: μ = k
H_a: μ ≠ k
α





$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$

We won't reject H_0 iff $-Z_{\alpha/2} < \frac{\overline{Y} - k}{\sigma/\sqrt{n}} < Z_{\alpha/2}$

algebra

$$\overline{Y} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < k < \overline{Y} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Therefore…
We won't reject H_0 iff the (1 – α)100% CI for μ captures k.
We will reject H_0 iff the (1 – α)100% CI for μ doesn't capture k.

*α must be the same
*same data
*H_a must be two-tailed

A 95% CI is (160,180)

Would we reject $\left\{\begin{matrix} H_0: \mu = 173 \\ H_a: \mu \neq 173 \\ \alpha = .05 \end{matrix}\right\}$?

No, since the 95% CI contains 173.

Would we reject $\left\{\begin{matrix} H_0: \mu = 175 \\ H_a: \mu \neq 175 \\ \alpha = .01 \end{matrix}\right\}$?

No, since the 99% CI would contain 175.

A 95% CI is (160,180)

Would we reject $\left\{\begin{matrix} H_0: \mu = 183 \\ H_a: \mu \neq 183 \\ \alpha = .05 \end{matrix}\right\}$?

Yes, since the 95% CI doesn't contain 183.

Would we reject $\left\{\begin{matrix} H_0: \mu = 183 \\ H_a: \mu \neq 183 \\ \alpha = .01 \end{matrix}\right\}$?

We don't know – it depends on whether or not the 99% CI contains 183.

Two things in the book we are skipping over:

*power & sample size
p. 197-202

*prediction intervals
p. 216

Proportions

p = proportion of people or things in the population with a certain characteristic

population

$0 \leq p \leq 1$

random

sample

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

p = proportion of students who walk to school = .27

n = 100
Y = 25

$\hat{p} = \frac{25}{100} = .25$

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

n = sample size

Y = # of people/things in sample with characteristic

Y = { 0, 1, 2, … n }

Y is Binomial
n trials, P(success) = p

$\mu = np, \sigma^2 = npq, \sigma = \sqrt{npq}$

<u>before</u>: $\overline{Y}$ estimates μ

<u>checklist</u>:

$\overline{Y}$ normal or approx. normal

$E\left(\overline{Y}\right)=\mu$ (unbiased)

$$SD\left(\overline{Y}\right)=\frac{\sigma}{\sqrt{n}}$$

<u>now</u>: $\hat{p}$ estimates p

<u>checklist</u>:

$\hat{p}$ approx. normal (must have $n \geq 30$)

$$E\left(\hat{p}\right)=E\left(\frac{Y}{n}\right)=\frac{1}{n}E\left(Y\right)=\frac{1}{n}\left(np\right)=p$$

(unbiased)

$$Var\left(\hat{p}\right)=Var\left(\frac{Y}{n}\right)=\frac{1}{n^2}Var\left(Y\right)=\frac{\sigma^2}{n^2}=\frac{npq}{n^2}=\frac{pq}{n}$$

$$SD\left(\hat{p}\right)=\sqrt{\frac{pq}{n}}$$

Confidence Interval Recipe:

Estimator ± (Z or T) SD(Estimator)

(1−α)100 % CI for p

$$\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Filling one gallon milk jugs. What proportion of jugs does our filling process successfully place at least one gallon of milk in?

n = 100 Y = 80 $\hat{p}=\frac{80}{100}=.80$

95% CI for p

$$\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}=.80\pm 1.96\sqrt{\frac{.8(.2)}{100}}$$

$$=.80\pm .0784=\boxed{(.7216,.8784)}$$

If the sample had been 800 out of 1000, the 95% CI would have been <u>thinner</u>.

$$.80\pm 1.96\sqrt{\frac{.8(.2)}{1000}}=.80\pm .0248=\boxed{(.7752,.8248)}$$

If we want the error margin to equal some small value B, what sample size is necessary?

$$B = Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \rightarrow \boxed{n = \hat{p}\hat{q}\left(\frac{Z_{\alpha/2}}{B}\right)^2}$$

If you have no prior estimate of p, use $\hat{p} = .50$

Milk jugs again. What if we wanted a 99% CI for p to have an error margin of .01? What should n be?

$\hat{p} = .80$

$$n = \hat{p}\hat{q}\left(\frac{Z_{\alpha/2}}{B}\right)^2 = .8(.2)\left(\frac{2.576}{.01}\right)^2 = 10617.2 \uparrow \boxed{10618}$$

If we had no prior estimate of p:

$$n = \hat{p}\hat{q}\left(\frac{Z_{\alpha/2}}{B}\right)^2 = .5(.5)\left(\frac{2.576}{.01}\right)^2 = 16589.4 \uparrow \boxed{16590}$$

Test Statistic Recipe:

(Z or T) = (estimator) – (hypothesized parameter value) / SD(estimator)

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

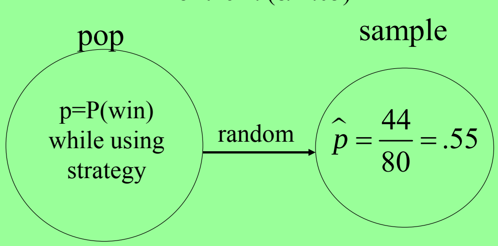
$$H_0 : p = p_0$$

Three options for H_a

$$\begin{Bmatrix} H_a : p < p_0 \\ H_a : p > p_0 \\ H_a : p \neq p_0 \end{Bmatrix}$$

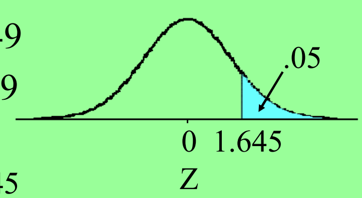
$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

A gambler has invented a new strategy for a game where players are supposed to win 49% of the time. He hopes his strategy will win more often. To test it, he plays the game 80 times and wins 44 of them. (α =.05)



$H_0 : p = .49$

$H_a : p > .49$



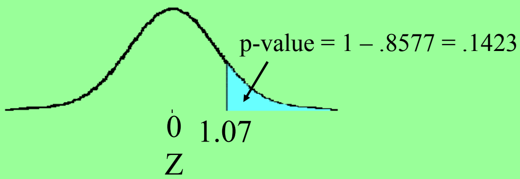
R: Z > 1.645

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.55 - .49}{\sqrt{\frac{.49(.51)}{80}}} = 1.07$$

don't reject $H_0 \rightarrow$ p-value > .05

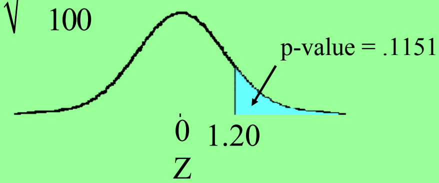
Data not significant at α =.05

The gambler does not have significant evidence that his strategy will win more than 49% of the time.



If the gambler had won 55 out of 100 games, the p-value would have been ___smaller___.

$$Z = \frac{.55 - .49}{\sqrt{\frac{.49(.51)}{100}}} = 1.20$$



$$\hat{p} = .55$$

Y	n	Z	p-value
44	80	1.07	.1423
55	100	1.20	.1151
275	500	2.68	.0037
550	1000	3.80	$\approx \frac{1}{14000}$

Section 4.7

A quick look at a hypothesis test for σ^2

$$E\left(s^2\right) = \sigma^2 \quad \text{(unbiased)}$$

What kind of r.v. is s^2? Not normal!

before:

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\overline{Y} - \mu}{s/\sqrt{n}} \qquad Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

now: If Y is normal, then

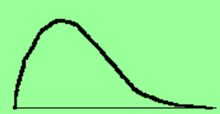
$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

Chi-Square Distribution χ^2_k k = d.o.f.

$$f(y) = \begin{cases} \dfrac{x^{\left(k/2 - 1\right)} e^{-\frac{x}{2}}}{2^{k/2}\Gamma\left(k/2\right)} & 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

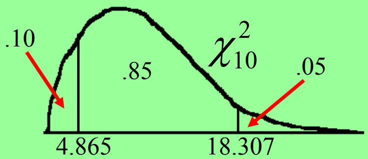
$$E\left(\chi^2_k\right) = k$$

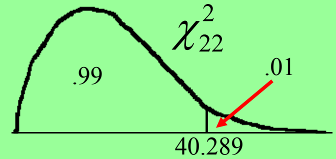
$$Var\left(\chi^2_k\right) = 2k$$

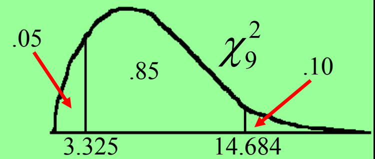


Chi-Square table

	10%	95%
10	4.865	18.307







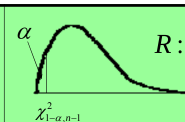
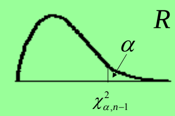
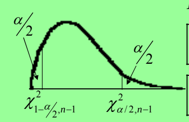
Hypothesis test for σ^2

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

Three options for H_a

$$\begin{Bmatrix} H_a : \sigma^2 < \sigma_0^2 \\ H_a : \sigma^2 > \sigma_0^2 \\ H_a : \sigma^2 \neq \sigma_0^2 \end{Bmatrix}$$

Test statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

$H_a : \sigma^2 < \sigma_0^2$ left-sided		$R: \chi^2 < \chi^2_{1-\alpha,n-1}$
$H_a : \sigma^2 > \sigma_0^2$ right-sided		$R: \chi^2 > \chi^2_{\alpha,n-1}$
$H_a : \sigma^2 \neq \sigma_0^2$ two-sided		R: $\boxed{\chi^2 < \chi^2_{1-\alpha/2,n-1}}$ OR $\boxed{\chi^2 > \chi^2_{\alpha/2,n-1}}$

Manufacturing sticks of dynamite – waiting time (seconds) for detonation is designed to be very close to 25, and obviously the variance should be very small.

The manufacturer wants $\begin{Bmatrix} \sigma^2 = 0.25\,\text{sec} \\ \sigma = 0.50\,\text{sec} \end{Bmatrix}$

and is afraid that it might be too big.

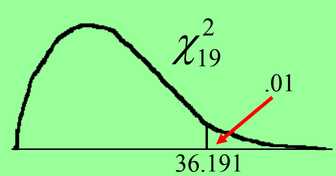
$$H_0 : \sigma^2 = 0.25$$
$$H_a : \sigma^2 > 0.25$$
$$\alpha = .01$$

n = 20

$\overline{Y} = 25$

$s^2 = 0.2916$

$s = .5400$



$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$R: \chi^2 > 36.191$$

$H_0 : \sigma^2 = 0.25$

$H_a : \sigma^2 > 0.25$

$\alpha = .01$

$R : \chi^2 > 36.191$

n = 20

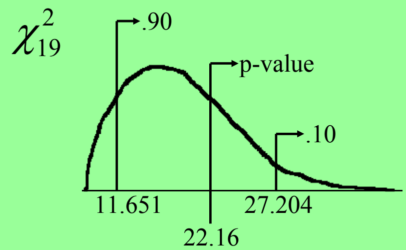
$\overline{Y} = 25$

$s^2 = 0.2916$

$s = .5400$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(19)(0.2916)}{0.25} = 22.16$$

don't reject $H_0 \rightarrow$ p-value > .01



.10 < p-value < .90

MA 3710

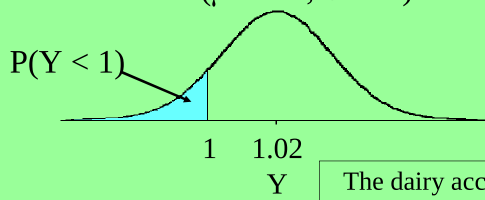
CH 5

Stonehenge
England

A dairy fills plastic milk jugs

Y = gallons of milk placed in a plastic jug

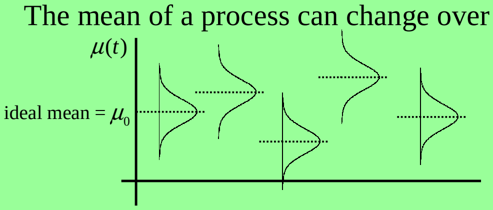
Y ~ N(μ=1.02, σ=.01)



$$P(Y<1)=.0228\approx\frac{1}{44}$$

The dairy accepts that about one in 44 jugs on average will be underfilled.

The mean of a process can change over time



$$Y_i=\mu(t)+\varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma)$$

*i*th observation

mean at time t

error

typical assumption

An engineer monitors a process with a <u>control chart</u>. He/she wants to know if/when any change occurs.

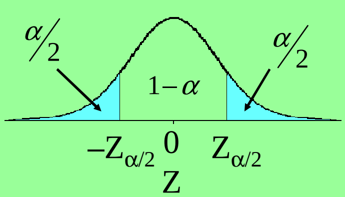
<u>Idea</u>: Periodically sample the process and conduct hypothesis tests with the data.

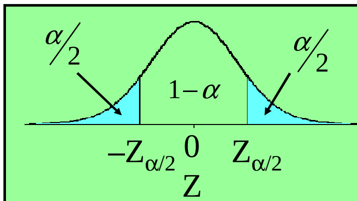
Two-tailed test for μ:

$H_0:\mu=\mu_0$

$H_a:\mu\neq\mu_0$

α





$$Z=\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$$

We won't reject H_0 iff $-Z_{\alpha/2}<\dfrac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}<Z_{\alpha/2}$

algebra

$$\boxed{\mu_0-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}<\bar{Y}<\boxed{\mu_0+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

$$\boxed{\mu_0 - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}} < \overline{Y} < \boxed{\mu_0 + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

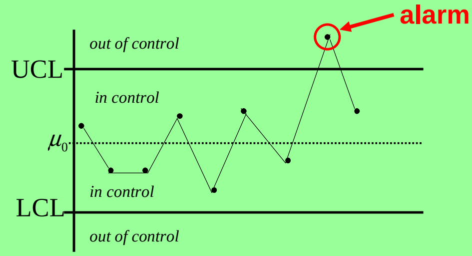
limits

lower control limit = LCL = $\mu_0 - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

upper control limit = UCL = $\mu_0 + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Take random samples at fixed time intervals and plot $\overline{Y}$

Each sample is a hypothesis test.



Tradition: $Z_{\alpha/2} = 3 \quad \alpha = .0027$

LCL = $\mu_0 - 3\frac{\sigma}{\sqrt{n}}$ UCL = $\mu_0 + 3\frac{\sigma}{\sqrt{n}}$

$\alpha = .05$ would be too high (too many false alarms)

Remember: even a process that is in control will false alarm about once in every $\frac{1}{\alpha} = \frac{1}{.0027} \approx 370$ samples

A dairy fills plastic milk jugs. They decide to take a sample of size n = 4 once a day and compute $\overline{Y}$

known	unknown
$\mu(t) = 1.02$ at t = 0	$\mu(t)$ when t > 0
$\sigma = .01$	
population is normal	
$Y \sim N(\mu(t), \sigma = .01)$	
$\overline{Y} \sim N(\mu(t), \sigma/\sqrt{n} = .01/\sqrt{4})$	

UCL = $\mu_0 + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.02 + 3\left(\frac{.01}{\sqrt{4}}\right) = 1.035$

LCL = $\mu_0 - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.02 - 3\left(\frac{.01}{\sqrt{4}}\right) = 1.005$

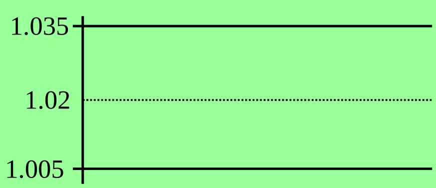


Chart A: Process in control for all 20 days

$\mu(t) = 1.02$ if $1 \le t \le 20$

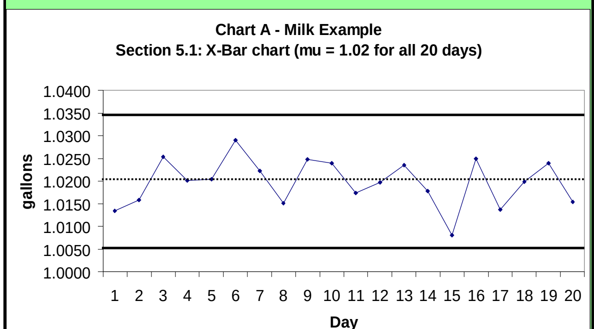
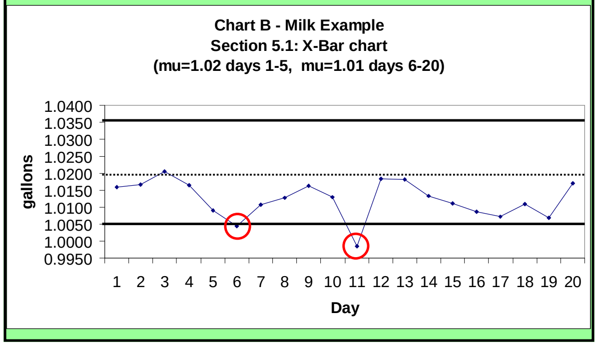


Chart B: Process out of control after Day 5

$\mu(t) = 1.02$ if $1 \le t \le 5$ $\mu(t) = 1.01$ if $6 \le t \le 20$



In real life, we usually don't know μ(t) (even at t = 0) or σ

Phase I chart
estimate parameters and compute control limits with the estimates

Phase II chart
monitor process

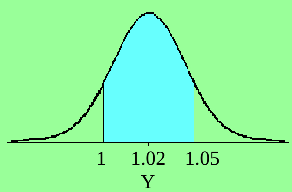
must assume: process is in control during Phase I

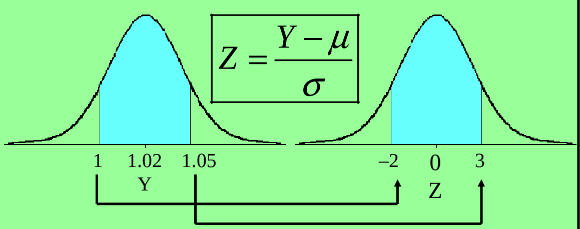
Control limits (UCL, LCL) are the "voice of the process", usually come from sample data, and are used to assess whether or not the process is stable.

Specification limits (USL, LSL) are the "voice of the customer" and define an acceptable range for the product.

A dairy fills plastic milk jugs. Suppose regulations require jugs to contain between 1 and 1.05 gallons.

If process is in control, find P(1 < Y < 1.05)
$Y \sim N(\mu = 1.02, \sigma = .01)$





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

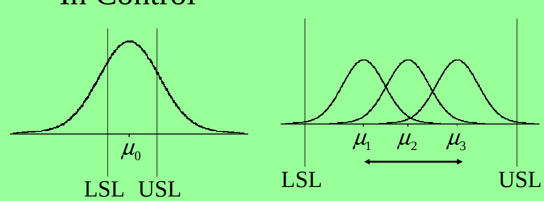
P(1<Y<1.05) =

$$P\left(\frac{1 - 1.02}{.01} < Z < \frac{1.05 - 1.02}{.01}\right)$$

= P(–2 < Z < 3) = .9987 – .0228 = $\boxed{.9759}$

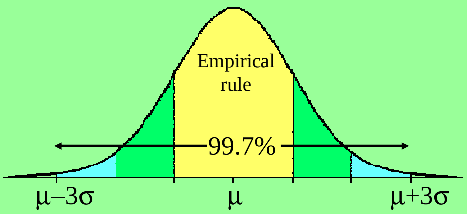
Note: it is possible for an in-control process to produce regular observations outside of specification limits and vice-versa.

In Control   Out of Control



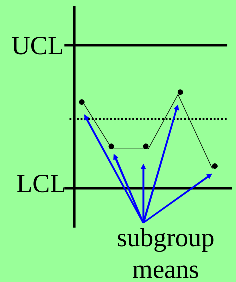
Capability index: $\frac{USL-LSL}{6\sigma}$

It's good if $\frac{USL-LSL}{6\sigma} > 1 \rightarrow USL-LSL > 6\sigma$



~~5.3 $\bar{X}$ and R chart~~ skip

$\boxed{5.4}$ $\bar{X}$ and s^2 chart



An $\bar{X}$ chart monitors <u>between-subgroup</u> variability and assumes that σ^2 (the <u>within-group</u> variance) is constant.

<u>Common sense</u>:
If σ^2 is changing over time, the $\bar{X}$ chart is meaningless

σ^2 should be monitored with its own chart

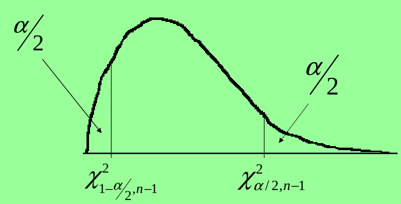
<u>Realistic situation</u>:
μ and σ^2 are unknown and are estimated during the Phase I "base period"

Two-sided test for σ^2

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

$H_a : \sigma^2 \neq \sigma_0^2$

α



Don't reject H_0 if: test stat

$$\chi^2_{n-1,1-\alpha/2} < \chi^2 < \chi^2_{n-1,\alpha/2}$$

$$\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2}$$

$$\Big\Downarrow \text{algebra}$$

$$\boxed{\frac{\left(\chi^2_{n-1,1-\alpha/2}\right)\sigma_0^2}{n-1}} < s^2 < \boxed{\frac{\left(\chi^2_{n-1,\alpha/2}\right)\sigma_0^2}{n-1}}$$

σ^2 must be estimated with $\overline{s^2}$ the average sample variance during the Phase I "base period" using traditional $\alpha = .0027 \rightarrow \alpha/2 = .00135$

$\overline{s^2}$ chart control limits:

$$\boxed{UCL = \frac{\left(\chi^2_{n-1,.00135}\right)\overline{s^2}}{n-1}} \qquad \boxed{LCL = \frac{\left(\chi^2_{n-1,.99865}\right)\overline{s^2}}{n-1}}$$

$\overline{X}$ chart control limits:

$$\boxed{UCL = \overline{\overline{Y}} + 3\sqrt{\frac{\overline{s^2}}{n}}} \qquad \boxed{LCL = \overline{\overline{Y}} - 3\sqrt{\frac{\overline{s^2}}{n}}}$$

$\overline{\overline{Y}}$ = average mean during Phase I base period

<u>assumptions</u>:

*process variance is constant

(s^2 chart should *always* be in control)

*subgroup means are in control during Phase I base period

*subgroup means are normal or approx. normal

(normal population or large samples)

Charts C & D

Manufacturing bolts, measuring shear strength (kip). A sample of n = 6 is taken every hour. 40 samples – first 20 are the Phase I base period.

$$\overline{\overline{Y}} = 4.999 \qquad \overline{s^2} = 2.291$$

$\overline{X}$ chart control limits:

$$\overline{\overline{Y}} \pm 3\sqrt{\frac{\overline{s^2}}{n}} = 4.999 \pm 3\sqrt{\frac{2.291}{6}} = \underset{\text{LCL}}{\boxed{3.145}} \text{ and } \underset{\text{UCL}}{\boxed{6.853}}$$

$\overline{s^2}$ chart control limits:

$$\frac{\left(\chi^2_{n-1,.00135}\right)\overline{s^2}}{n-1} = \frac{19.821(2.291)}{5} = \underset{\text{UCL}}{\boxed{9.083}}$$

$$\frac{\left(\chi^2_{n-1,.99865}\right)\overline{s^2}}{n-1} = \frac{0.238(2.291)}{5} = \underset{\text{LCL}}{\boxed{0.109}}$$

Chart C: in control until hour 39

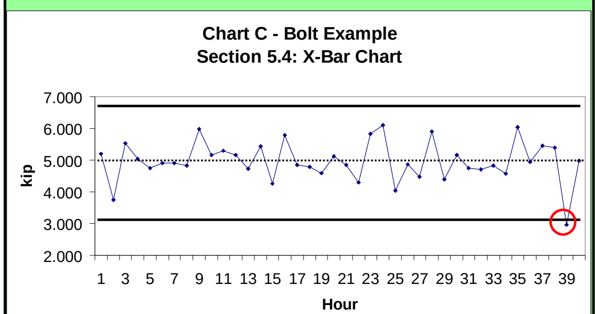
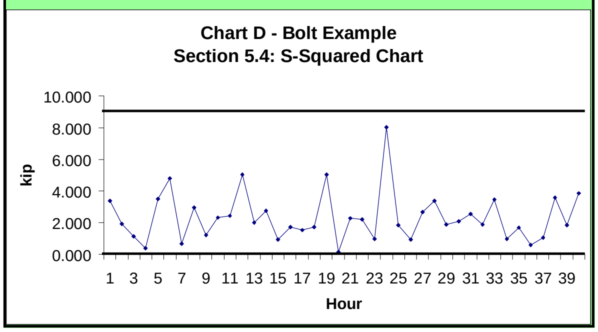


Chart D: in control



$\boxed{5.5}$ X chart

Some processes do not produce data frequently enough to have regular samples of n > 1, so we need to use individual observations.

Normal data:

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

$$\bar{Y} \sim N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$$

control limits:

before: $\mu_0 \pm 3\frac{\sigma}{\sqrt{n}}$ now?: $\cancel{\mu_0 \pm 3\sigma}$ No.

Realistically, μ and σ are unknown. For this chart, we will investigate process variability with the moving range = MR.

Observations: $Y_1, Y_2, Y_3, Y_4, \ldots$

Base period: first m observations

$$MR_i = |Y_{i+1} - Y_i|$$

$$\overline{MR} = \frac{1}{m-1}\sum_{i=1}^{m-1} MR_i$$

i	Y_i	MR_i
1	7	5
2	2	0
3	2	3
4	5	2
5	3	7
6	10	

m=6

$$\overline{MR} = \frac{1}{5}(5+0+3+2+7) = \frac{17}{5} = \boxed{3.4}$$

control limits:

$$\mu_0 \pm 3\frac{\overline{MR}}{d_2}$$

estimated with $\bar{Y}$

1.128 (n=2)

$$\boxed{UCL = \bar{Y} + 2.66\overline{MR}} \quad \boxed{LCL = \bar{Y} - 2.66\overline{MR}}$$

Assumptions:
Data is normally distributed
Process is in control during base period

Chart E

In an underdeveloped country, a polluted river near an industrial plant is tested monthly for mercury contamination (Hg ppm). A second plant will be built in two years and scientists are concerned this may pollute the river even further.

Base Period: 24 months before new plant begins operation

Chart runs for 48 months total.

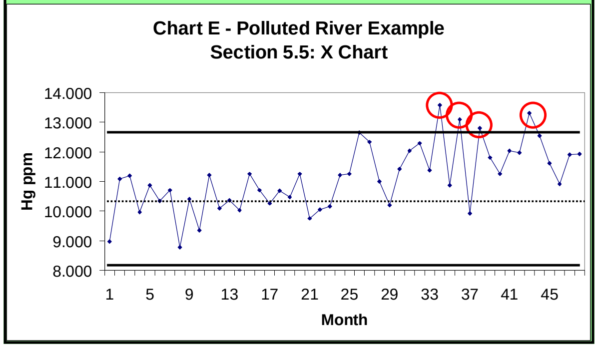
base period: $m = 24$, $\bar{Y} = 10.382$, $\overline{MR} = 0.870$

$$UCL = \bar{Y} + 2.66\overline{MR} = 10.382 + 2.66(0.87) = \boxed{12.695}$$

$$LCL = \bar{Y} - 2.66\overline{MR} = 10.382 - 2.66(0.87) = \boxed{8.069}$$

Chart E: process seems out of control

four observations beyond upper limit at months 34, 36, 38, 43. The second plant has most likely increased the mean Hg ppm in the river.



$\boxed{5.6}$ np chart

For monitoring the number of defectives in samples of size n (constant).

Y_i = # of defectives in ith sample

<u>assume</u>: Y is binomial, P(defective) = p, n trials

$$E(Y_i) = np \qquad SD(Y_i) = \sqrt{npq}$$

$$Var(Y_i) = npq$$

$$H_0: p = p_0 \rightarrow np = np_0$$

$$H_a: p \neq p_0 \rightarrow np \neq np_0$$

Nice if $np_0 \geq 10$ and $nq_0 \geq 10$ so the LCL will be positive.

Limits: $np_0 \pm 3\sqrt{np_0 q_0}$

p_0 estimated with $\bar{p}$ = average proportion of defectives during base period

m = length of base period $\qquad \bar{p} = \dfrac{1}{mn}\sum_{i=1}^{m} Y_i$

$$\boxed{UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}} \qquad \boxed{LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}}$$

Chart F

Making batches of n=15,000 pencils per day: defectives are counted every day

total chart length = 365 days

base period = m = 31 (January (assume in control))

$$\bar{p} = \frac{1}{mn}\sum_{i=1}^{m} Y_i = \frac{1}{31(15000)}[474] = \boxed{.00102} \approx \frac{1}{981}$$

31 days

474 defectives in Jan.

15000 pencils per day

Limits: $n\bar{p} \pm 3\sqrt{n\bar{p}\bar{q}}$

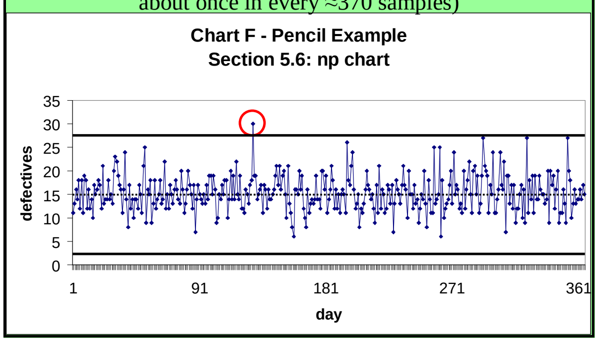
$$= 15000(.00102) \pm 3\sqrt{15000(.00102)(.99898)}$$

$$= 15.3 \pm 11.729$$

$$= \boxed{3.571} \text{ and } \boxed{27.029}$$

Chart F: process seems in control

one observation above upper limit on May 9 (probably not a concern since we expect an in-control process to false alarm about once in every ≈370 samples)



$\boxed{5.7}$ c chart

For monitoring the number of occurrences of an event per time period

Y_i = # of occurrences in ith period

assume: Y is Poisson, λ = mean rate of occurrences

$$E(Y_i) = \lambda \qquad SD(Y_i) = \sqrt{\lambda}$$

$$Var(Y_i) = \lambda$$

$$H_0 : \lambda = \lambda_0$$
$$H_a : \lambda \neq \lambda_0$$

Nice if $\lambda_0 \geq 10$ so the LCL will be positive.

Limits: $\lambda_0 \pm 3\sqrt{\lambda_0}$

λ_0 estimated with $\bar{c}$ = average count over base period

m = length of base period $\qquad \bar{c} = \frac{1}{m}\sum_{i=1}^{m} c_i$

$$\boxed{UCL = \bar{c} + 3\sqrt{\bar{c}}} \qquad \boxed{LCL = \bar{c} - 3\sqrt{\bar{c}}}$$

Chart G

Software developers are keeping track of how many calls per week the service department receives

total chart length = 100 weeks

base period = m = 20 (assume in control)

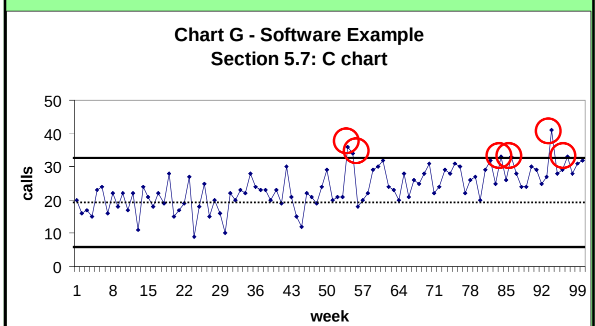
$$\bar{c} = \frac{1}{m}\sum_{i=1}^{m} c_i = \frac{1}{\underset{\text{20 weeks}}{20}}\underset{\text{390 calls during first 20 weeks}}{[390]} = \boxed{19.5}$$

Limits:

$$\bar{c} \pm 3\sqrt{\bar{c}} = 19.5 \pm 3\sqrt{19.5} = \boxed{6.252} \text{ and } \boxed{32.748}$$

Chart G: process seems to be out of control (chart seems to have an increasing trend)

six observations above upper limit during weeks 54, 55, 84, 86, 94, 97



$\boxed{5.8}$

N = # of samples until control chart signals

average run length = ARL = E(N) = $\frac{1}{\text{P(alarm)}}$

N is a geometric random variable:

Y = # of trials needed to observed first success

$$E(Y) = \frac{1}{p}$$

An in-control process using the traditional $Z_{\alpha/2} = 3$ and $\alpha = .0027$ has

$$ARL = \frac{1}{.0027} \approx 370$$

σ known to be 16. μ supposed to be $\mu_0 = 100$

If n = 4, limits are

$$\mu_0 \pm 3\frac{\sigma}{\sqrt{n}} = 100 \pm 3\frac{16}{\sqrt{4}} = 100 \pm 24 = \boxed{76} \text{ and } \boxed{124}$$

If n = 9, limits are

$$\mu_0 \pm 3\frac{\sigma}{\sqrt{n}} = 100 \pm 3\frac{16}{\sqrt{9}} = 100 \pm 16 = \boxed{84} \text{ and } \boxed{116}$$

if μ actually equals 116 (μ + σ), on average how long will we wait for the chart to signal?

n = 4

P(signal) = 1 – P(no signal)

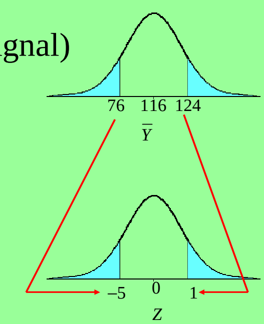
$$= 1 - P(76 < \overline{Y} < 124)$$

$$= 1 - P\left(\frac{76-116}{16/\sqrt{4}} < \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} < \frac{124-116}{16/\sqrt{4}}\right)$$

$$= 1 - P(-5 < Z < 1)$$

$$= 1 - [.8413 - (\approx 0)]$$

$$= \boxed{.1587}$$



n = 9

P(signal) = 1 – P(no signal)

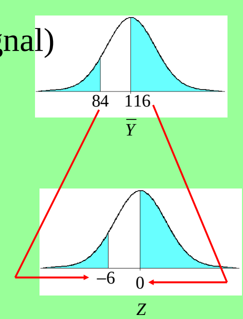
$$= 1 - P(84 < \overline{Y} < 116)$$

$$= 1 - P\left(\frac{84-116}{16/\sqrt{9}} < \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} < \frac{116-116}{16/\sqrt{9}}\right)$$

$$= 1 - P(-6 < Z < 0)$$

$$= 1 - [.5000 - (\approx 0)]$$

$$= \boxed{.5000}$$



$\mu_0 = 100$ and σ = 16

if n = 4

actual value of μ	P(signal)	ARL
100	.0027	370.37
116	.1587	6.30
132	.8413	1.19

if n = 9

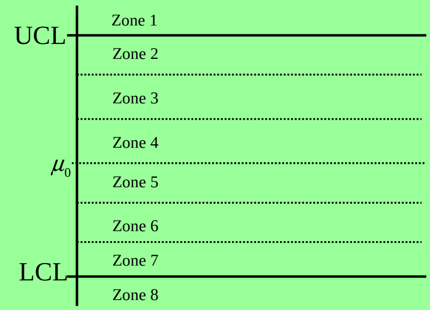
actual value of μ	P(signal)	ARL
100	.0027	370.37
116	.5000	2
132	≈ 1	≈ 1

ARL is smaller if:

*actual value of μ is further away from μ_0

*n is larger

5.9 Runs Rules



"Regular" Rule:

A chart signals if an observation lands in Zone 1 or Zone 8

Other Rules exist and may be applied

Trade-off:

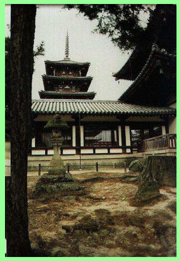
Good: more likely to detect an out of control process

Bad: more likely to false alarm

5.10 CUSUM and EWMA charts

MA 3710

CH 6



Horyu-Ji Temple
Nara, Japan

<u>Scatterplot</u>

paired data (x, y)

We would like to predict one variable with another.

population

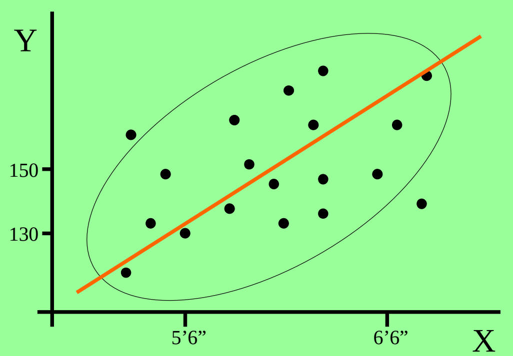
(x,y) 

X = height
Y = weight

sample

(x_1, y_1)
$(x_2 y_2)$
...
(x_n, y_n)

trend: taller people tend to be heavier



<u>association</u> – as X increases, Y tends to increase (or decrease)

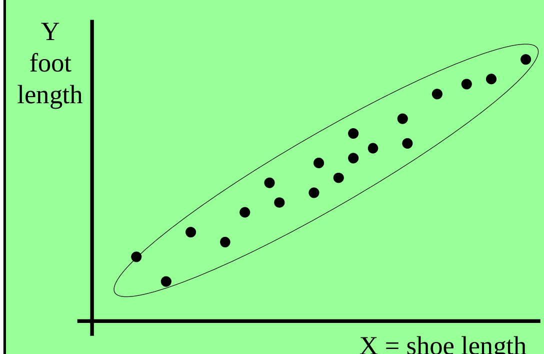
In MA 3710 we will only be interested in <u>linear</u> associations.

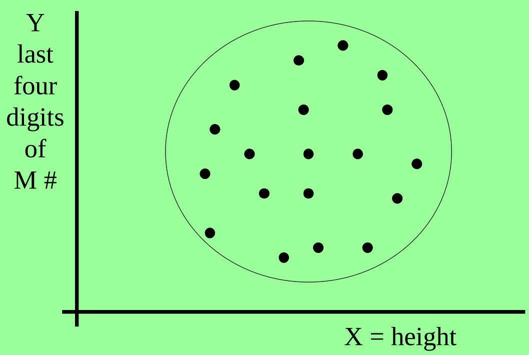
If there is a positive or negative association between X and Y, knowing an X value can help predict a Y value.

X = independent variable
(explanatory, predictor, regressor)

Y = dependent variable
(response)

pounds of fertilizer = X
pounds of soybeans harvested = Y





r = correlation coefficient

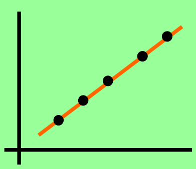
r measures the strength of linear association between X and Y

$$-1 \le r \le 1$$

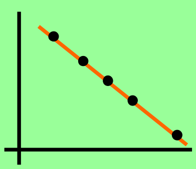
"strong" associations are close to 1 or –1

"weak" associations are close to zero.

r equals 1 or –1 if Y is a linear function of X



people
X = height in inches
Y = height in cm
r = 1



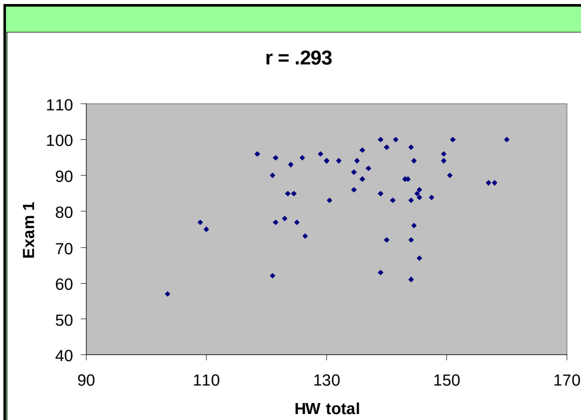
multiple-choice quiz
X = # correct
Y = # incorrect
r = –1

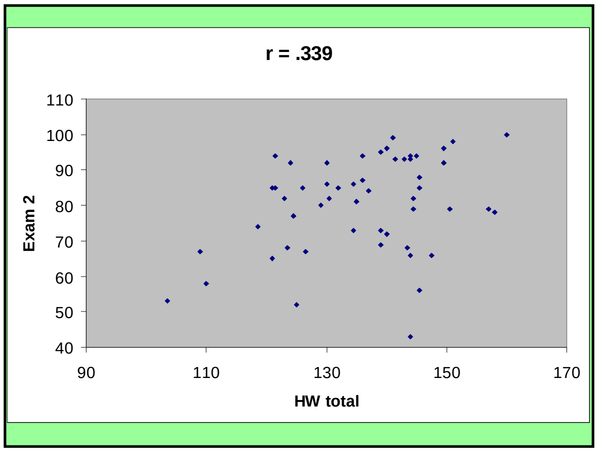
Correlation Matrix   STAT 301 data

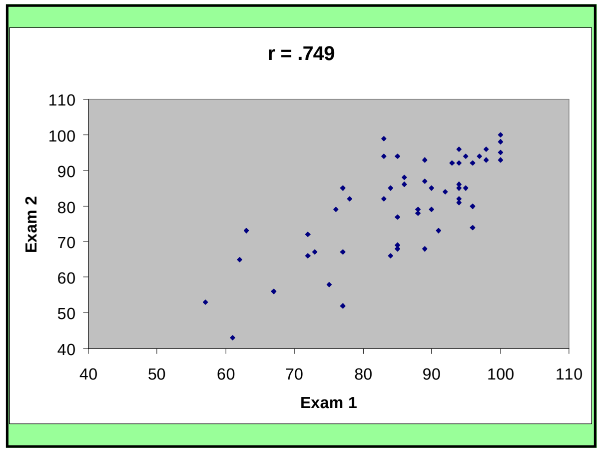
ID#	1			
Pop Quizzes	–.063	1		
Exam 1	–.146	.098	1	
Exam 2	–.023	.379	.492	1
	ID#	PQ	Ex1	Ex2

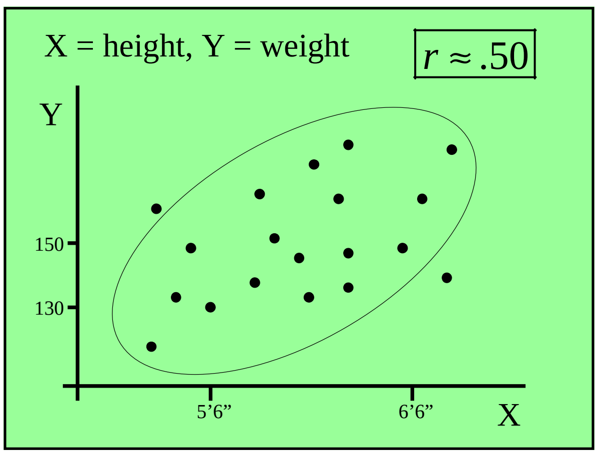
MA 3710 data
Fall 2008

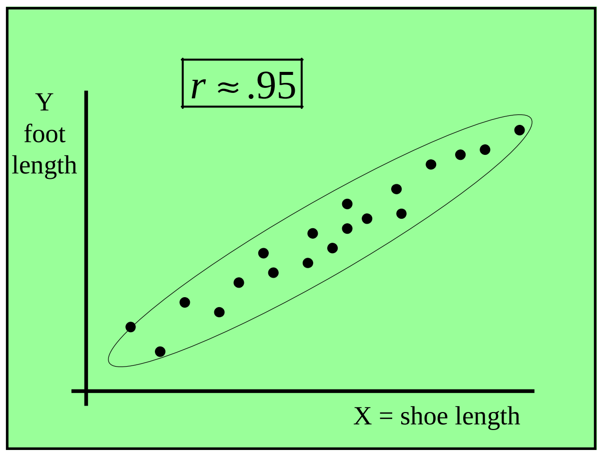
HW total (first 6)	1		
Exam 1	.293	1	
Exam 2	.339	.749	1
	HWt	Ex1	Ex2

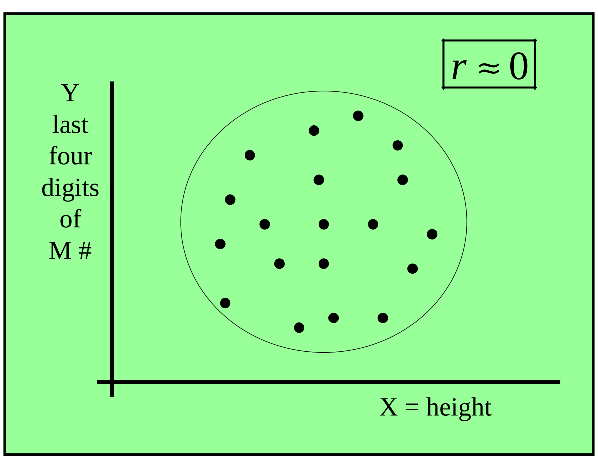












Computing r

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = (n-1)\,s_x^2$$

$$SS_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = (n-1)\,s_y^2$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

Computing r

$$\boxed{r = \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}}} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{(n-1)s_x s_y} = \frac{\sum Z_x Z_y}{n-1}$$

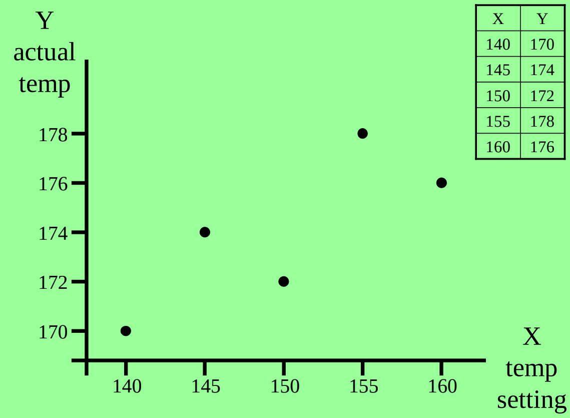
compute r

understand r

An engineering student has designed a new clothes dryer.
X = temperature setting °F
Y = actual interior temperature during cycle °F

n = 5

X	Y
140	170
145	174
150	172
155	178
160	176



X	Y
140	170
145	174
150	172
155	178
160	176

X	Y	X^2	Y^2	XY
140	170	19600	28900	23800
145	174	21025	30276	25230
150	172	22500	29584	25800
155	178	24025	31684	27590
160	176	25600	30976	28160
750	870	112750	151420	130580

$$\overline{X} = \frac{750}{5} = 150 \qquad \overline{Y} = \frac{870}{5} = 174$$

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 112750 - \frac{750^2}{5} = \boxed{250}$$

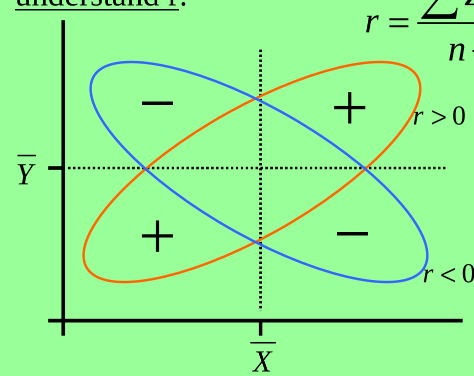
$$SS_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 151420 - \frac{870^2}{5} = \boxed{40}$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 130580 - \frac{750(870)}{5} = \boxed{80}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}} = \frac{80}{\sqrt{250}\sqrt{40}} = \boxed{.80}$$

understand r:

$$r = \frac{\sum Z_x Z_y}{n-1}$$



X = temperature setting °F
Y = actual interior temperature during cycle °F

r = .80

X = actual interior temperature during cycle °F
Y = temperature setting °F

r = .80

X = temperature setting °C
Y = actual interior temperature during cycle °C

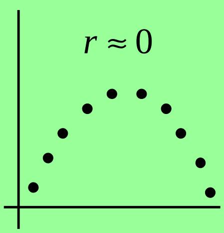
r = .80

The value of r doesn't change if X & Y are switched or if the units of measurement are changed.

Dangers with r
watch out for….

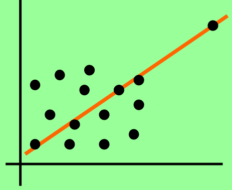
non-linearity

$r \approx 0$



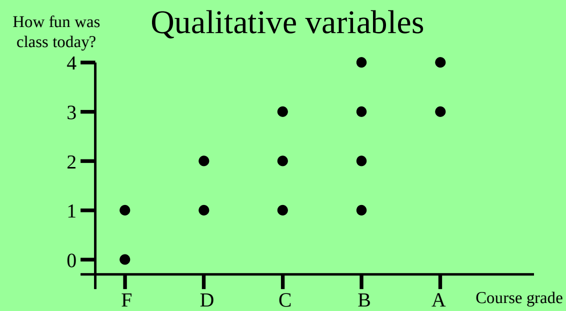
outliers

$r > 0$



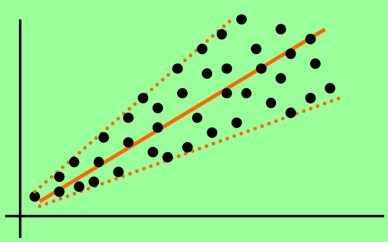
Dangers with r
watch out for….

Qualitative variables



Dangers with r
watch out for….

Variance that isn't constant



Cone shapes are bad

Dangers with r
watch out for….

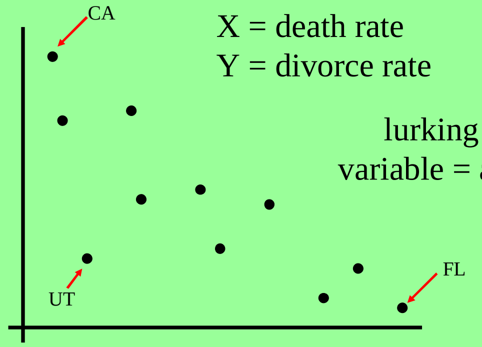
Association is not Causation

<u>association</u> – one variable tends to act along with the other

<u>causation</u> – one variable directly causes the other to behave in a certain manner

U.S. Census Data
X = death rate
Y = divorce rate

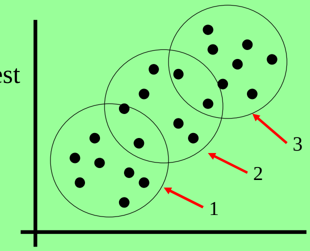
lurking variable = age



A <u>lurking variable</u> is not directly measured but is responsible for the association between two other variables.

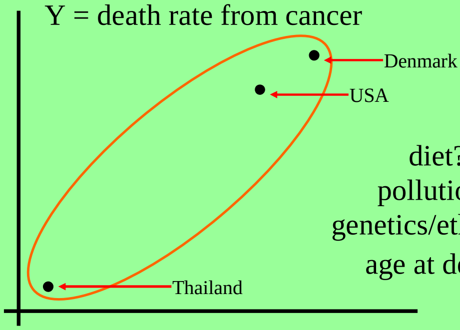
Elementary School
X = height
Y = score on reading test

lurking variable = grade



Nations of the world
X = fat intake per capita
Y = death rate from cancer

Mystery Correlation

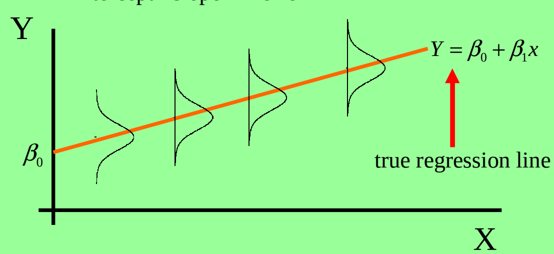


diet?
pollution?
genetics/ethnicity?
age at death?

If X and Y have a linear relationship, then

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

intercept slope error constant

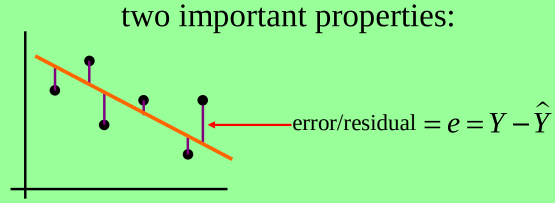


The true regression line is unknown (invisible). If a scatterplot appears to be linear with a constant variance, then we can estimate the true regression line with the <u>least-squares regression line</u>.

$$\hat{Y} = b_0 + b_1 x$$

<u>Unknown parameters</u>	<u>estimators</u>
β_0	$b_0 = \hat{\beta}_0$
β_1	$b_1 = \hat{\beta}_1$
σ^2	$s^2 = MS_{res}$

The <u>least-squares regression line</u> has two important properties:



$\sum e = 0$ The sum of the errors is zero.

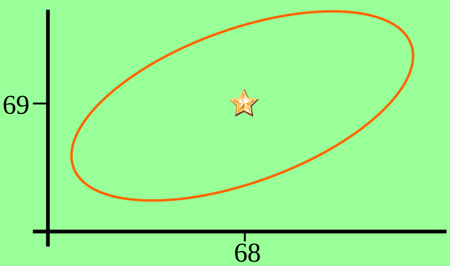
$\sum e^2 = SS_{res}$ is minimized. ("least-squares")

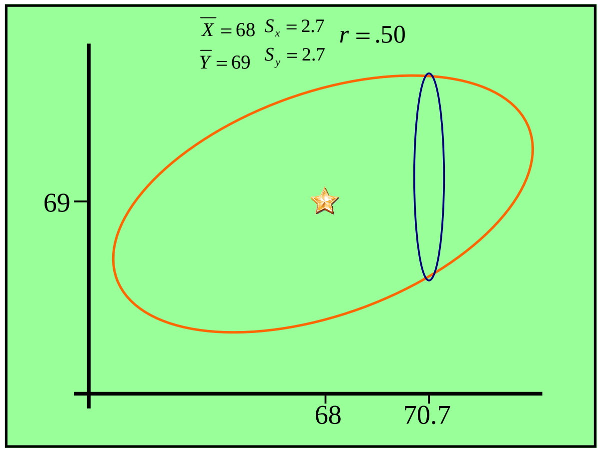
Francis Galton (1822-1911)

X = fathers height $\overline{X} = 68$ $S_x = 2.7$

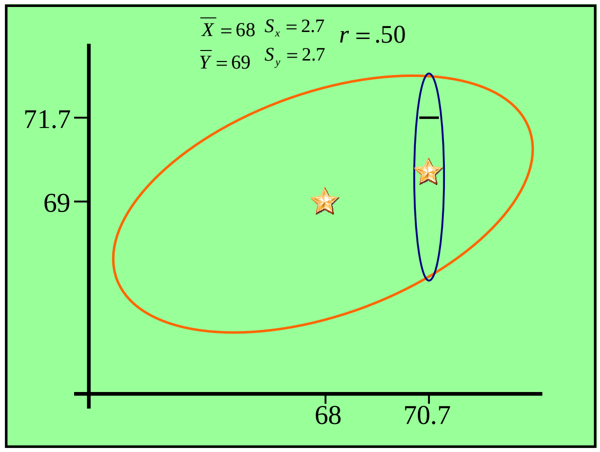
Y = sons height $\overline{Y} = 69$ $S_y = 2.7$

$r = .50$

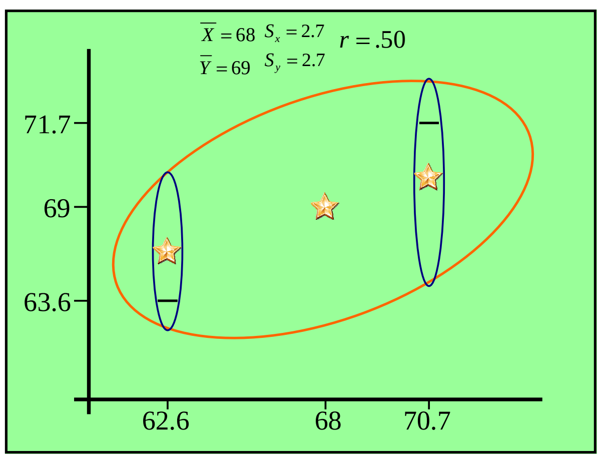




If a group of fathers are all 70.7 inches tall, (1 S_x above $\overline{X}$), what is our prediction for the mean height among their sons?

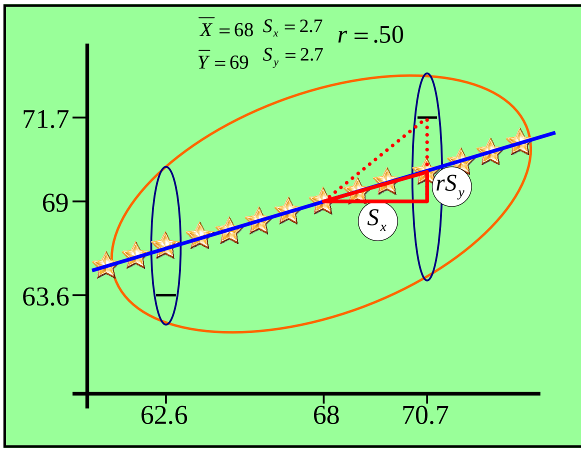


<u>Myth</u> – if an X value is 1 S_x above $\overline{X}$, then the corresponding Y values will on average be 1 S_y above $\overline{Y}$.



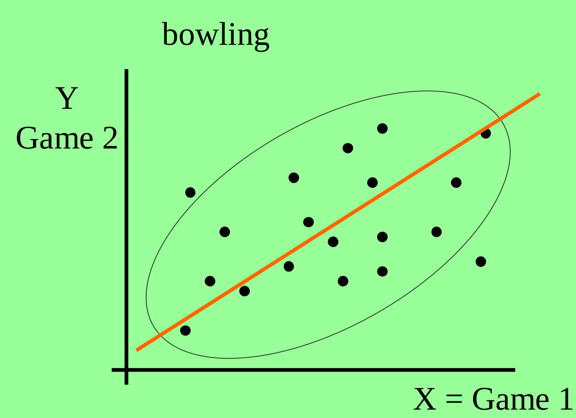
Galton – “observations tend to regress towards the mean”

Truth – if an X value is $1S_x$ above $\bar{X}$, then the corresponding Y values will on average be r S_y above $\bar{Y}$.



Truth – we are tempted to believe the slope of the least-squares line should be $\frac{S_y}{S_x}$, but instead it is $\boxed{\frac{rS_y}{S_x}}$

“Regression effect” – extreme observations usually aren’t as extreme the second time around



Least-Squares Regression Line

$$\hat{Y}=b_0+b_1X$$

Always goes through the point $(\bar{X},\bar{Y})$

$$b_1=\frac{rS_y}{S_x}=\frac{SS_{xy}}{SS_{xx}}$$

$$b_0=\bar{Y}-b_1\bar{X}$$

intuitive formula

MA 3710

Galton’s Data $\bar{X}=68$ $S_x=2.7$ $\bar{Y}=69$ $S_y=2.7$ $r=.50$

$$b_1=\frac{rS_y}{S_x}=\frac{.50(2.7)}{2.7}=.50$$

$$b_0=\bar{Y}-b_1\bar{X}=69-(.50)(68)=35$$

$$\boxed{\hat{Y}=35+.50X}$$

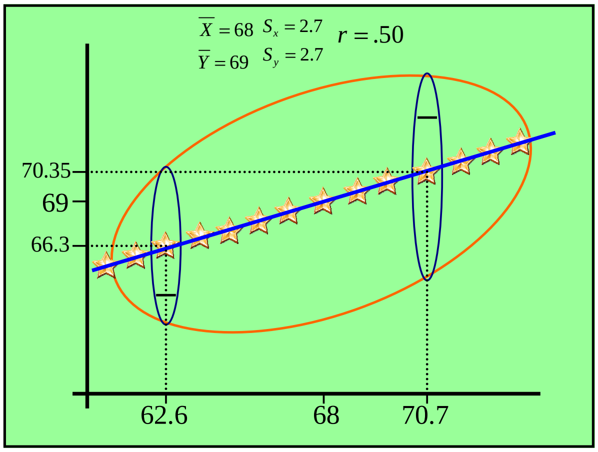
If X = 70.7, then

$$\hat{Y} = 35 + .50(70.7) = \boxed{70.35} \text{ inches}$$

If X = 62.6, then

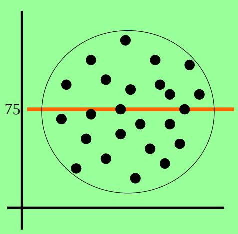
$$\hat{Y} = 35 + .50(62.6) = \boxed{66.30} \text{ inches}$$

*we predict average values, not specific observations



X = last 2 digits of M#
Y = exam score

$\bar{X} = 50$ $S_x = 20$
$\bar{Y} = 75$ $S_y = 15$
$r = 0$



$$b_1 = \frac{rS_y}{S_x} = \frac{0(15)}{20} = 0$$

$$b_0 = \bar{Y} - b_1\bar{X} = 75 - (0)(50) = 75$$

$$\boxed{\hat{Y} = 75 + 0X = 75}$$

An engineering student has designed a new clothes dryer.
X = temperature setting °F
Y = actual interior temperature during cycle °F

X	Y
140	170
145	174
150	172
155	178
160	176

$\bar{X} = 150$
$\bar{Y} = 174$
$S_x = 7.91$
$S_y = 3.16$

from previous work:
$SS_{xx} = 250$
$SS_{yy} = 40$
$SS_{xy} = 80$

$$b_1 = \frac{rS_y}{S_x} = \frac{SS_{xy}}{SS_{xx}} = \frac{80}{250} = 0.32$$

$$\frac{0.8(3.16)}{7.91} = 0.32$$

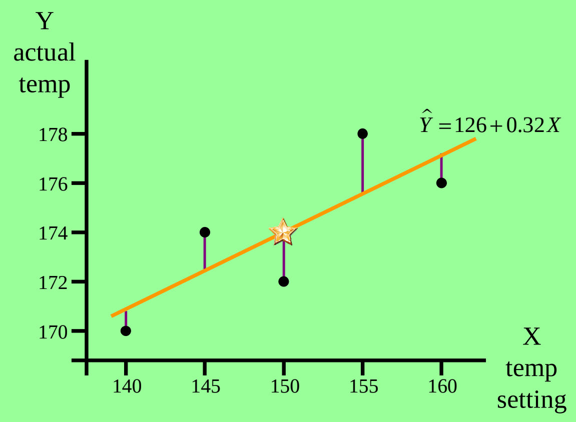
$$b_0 = \bar{Y} - b_1\bar{X} = 174 - (0.32)(150) = 126$$

$$\boxed{\hat{Y} = 126 + 0.32X}$$

$$\boxed{\hat{Y} = 126 + 0.32X}$$

On average, we expect a 0.32 degree increase in Y for every 1 degree increase in X

~~On average, we expect a temp of 126 °F if the dryer setting is 0 °F~~

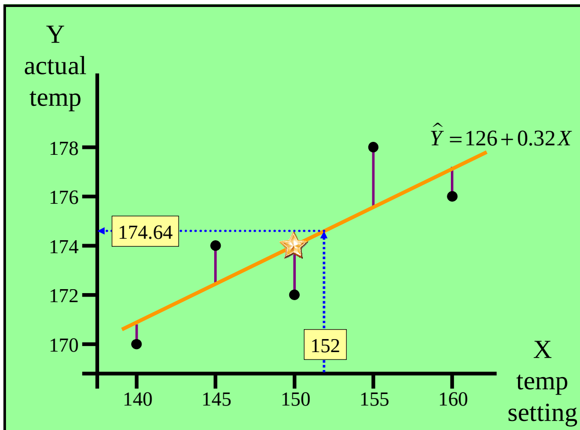


X	Y	$\hat{Y}$ $126+0.32X$	error residual $Y-\hat{Y}=e$	e^2
140	170	170.8	–0.8	0.64
145	174	172.4	1.6	2.56
150	172	174.0	–2.0	4.00
155	178	175.6	2.4	5.76
160	176	177.2	–1.2	1.44
$\sum e = 0$ always			0	14.40

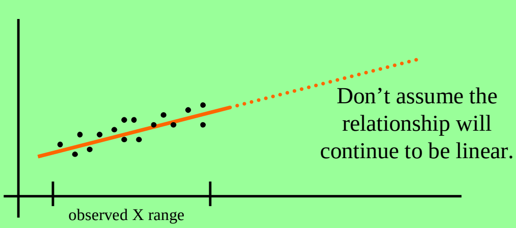
Any other line would have a larger total → $\sum e^2 = SS_{res}$

On average, we expect the dryer to run at 174.64 °F if we set it at X = 152 °F.

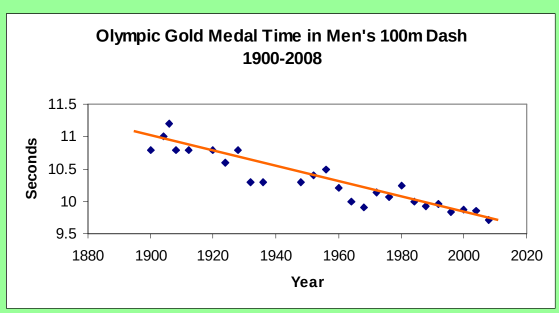
$$\hat{Y} = 126 + 0.32(152) = 174.64$$



*don't <u>extrapolate</u>! Only predict $\hat{Y}$ within your observed range of X values.



X = year
Y = gold medal time in men's 100 m dash



Evaluating a Linear Model

There is variation in Y. How much is due to X (the explanatory variable) and how much is due to chance (σ^2)?

Total variation in Y = Variation in X + Variation due to error

---

Total variation in Y = Variation in X + Variation due to error

$$SS_{total} = SS_{reg} + SS_{res}$$

$$\boxed{SS_{total}} = \sum\left(Y - \bar{Y}\right)^2 = \boxed{SS_{yy}}$$

$$\boxed{SS_{reg}} = \sum\left(\hat{Y} - \bar{Y}\right)^2 = \boxed{\frac{\left(SS_{xy}\right)^2}{SS_{xx}}}$$

$$\boxed{SS_{res}} = \sum\left(Y - \hat{Y}\right)^2 = \sum e^2 = \boxed{SS_{total} - SS_{reg}}$$

---

r^2 = coefficient of determination

$$0 \le r^2 \le 1$$

r^2 = the proportion of variation in Y that has been explained by X

$$r^2 = \frac{SS_{reg}}{SS_{total}}$$

The bigger r^2 is, the better.

Dryer example:

<u>from previous work</u>:

$$SS_{xx} = 250$$

$$SS_{yy} = 40$$

$$SS_{xy} = 80$$

$$r = .80$$

$$SS_{total} = SS_{yy} = \boxed{40}$$

$$SS_{reg} = \frac{\left(SS_{xy}\right)^2}{SS_{xx}} = \frac{80^2}{250} = \boxed{25.6}$$

$$SS_{res} = SS_{total} - SS_{reg} = 40 - 25.6 = \boxed{14.4}$$

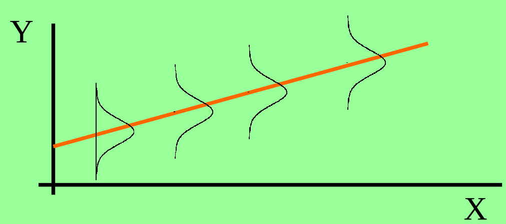
$$r^2 = \frac{SS_{reg}}{SS_{total}} = \frac{25.6}{40} = .64$$

$$r^2 = (.80)^2 = .64$$

"64% of the variation in dryer temp (Y) is explained by temp setting (X)."

True regression line: $Y = \beta_0 + \beta_1 x + \varepsilon$

$\varepsilon \sim N(0, \sigma)$ Determines width of bell curve



Estimate σ^2 with $s^2 = MS_{res} = \dfrac{SS_{res}}{n-2}$

Dryer example:

from previous work:

$SS_{res} = 14.4$

$n = 5$

$$s^2 = MS_{res} = \frac{SS_{res}}{n-2} = \frac{14.4}{3} = \boxed{4.8}$$

Testing the slope

Hypothesis test for β_1

$$H_0 : \beta_1 = 0$$

(X and Y are independent: X is a useless predictor of Y)

$$H_a : \beta_1 \neq 0$$

(X and Y are dependent)

Testing the slope

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

$$T = \frac{b_1 - 0}{\sqrt{\frac{MS_{res}}{SS_{xx}}}} = \frac{b_1}{\sqrt{\frac{MS_{res}}{SS_{xx}}}}$$

On a t-curve with n – 2 d.o.f.

Dryer example:

from previous work:

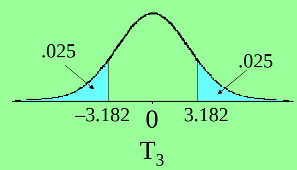
$n = 5$

$SS_{xx} = 250$

$b_1 = 0.32$

$MS_{res} = 4.8$

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$
$$\alpha = .05$$



R: T < –3.182 or T > 3.182

R: T < –3.182 or T > 3.182

$$T = \frac{b_1}{\sqrt{\frac{MS_{res}}{SS_{xx}}}} = \frac{0.32}{\sqrt{\frac{4.8}{250}}} = 2.31$$

*don't reject H_0

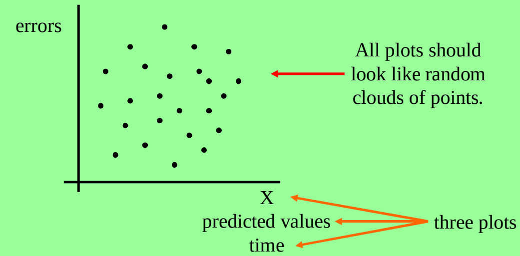
*data not significant at $\alpha = .05$

Computer Output

ANOVA = Analysis Of Variance
Simple Linear Regression

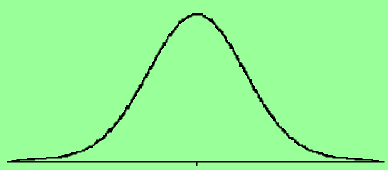
Source	d.o.f.	SS	MS	F
Regression	1	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{1}$	$F = \frac{MS_{reg}}{MS_{res}}$
Error	n – 2	SS_{res}	$MS_{res} = \frac{SS_{res}}{n-2}$	
Total	n – 1	SS_{total}		

*Residual/Error Plots can reveal problems with your data you hadn't seen before.



*Since we assume $\varepsilon \sim N(0,\sigma)$

A histogram of the errors should be bell-shaped.



A quick look at multiple regression

A university professor always gives Exam 2 on March 18. He is interested in predicting the scores. Randomly selects n = 25 students.

*Simple Linear Regression

Y = Exam 2

X = Exam 1

$$r^2 = .399 \approx .40$$

$$\hat{Y} = 7.07 + 0.87X$$

Only 40% of the variation in Exam 2 is explained by Exam 1. Other variables exist – wouldn't it be nice to include them in the model?

multiple regression:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

$$\hat{Y} = b_0 + b_1 x_1 + \ldots + b_k x_k$$

X_1 = Exam 1 $[0,100]$

X_2 = hours studied $[0,\infty)$

X_3 = hours of sleep $[0,\infty)$

X_4 = GPA $[0,4]$

X_5 = ACT $[0,36]$

X_6 = attendance $[0,20]$

X_7 = beers drank $[0,\infty)$

"Find the simplest model that works the best." –Prof. Tom Drummer

Warning: r^2 will increase if variables are added, regardless of what those variables are.

If (# of variables) = n – 1, then $r^2 = 1$

Final Model:

$$\hat{Y} = -43.141 + 0.505x_1 + 2.22x_2 + 2.79x_3$$

Exam 1 sleep ACT

MA 3710
CH 7

Assembly Hall
Salt Lake City,
Utah

$\boxed{7.1}$ 2^2 Factorial Experiment Design

Remote control car competition: Jim & Rachel design an experiment and want their car to run the course in the shortest time.

response variable Y = time to complete course (seconds)

Factor A: Octane
Levels: 90 or 95

"low" $x_1 = -1$ "high" $x_1 = 1$

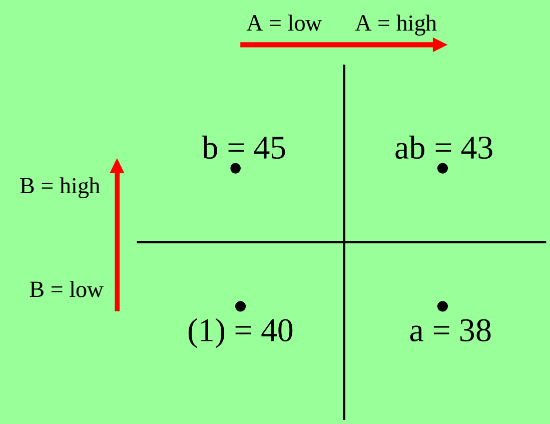
Factor B: Operator
Levels: Jim or Rachel

"low" $x_2 = -1$ "high" $x_2 = 1$

<u>treatment combinations</u>:

X_1 90 or 95	X_2 Jim or Rachel	description	average among n = 3 runs
–1	–1	90/Jim	(1) = 40
1	–1	95/Jim	a = 38
–1	1	90/Rachel	b = 45
1	1	95/Rachel	ab = 43

run in random order



	Factor A		
	X_1 = –1 90	X_1 = 1 95	
Factor B: X_2 = 1 Rachel	b = 45	ab = 43	$\boxed{44}$
Factor B: X_2 = –1 Jim	(1) = 40	a = 38	$\boxed{39}$
	$\boxed{42.5}$	$\boxed{40.5}$	$\boxed{41.5}$

---

effect of A
(from –1 to 1)
(90 to 95)

$$= \frac{a+ab}{2} - \frac{(1)+b}{2}$$

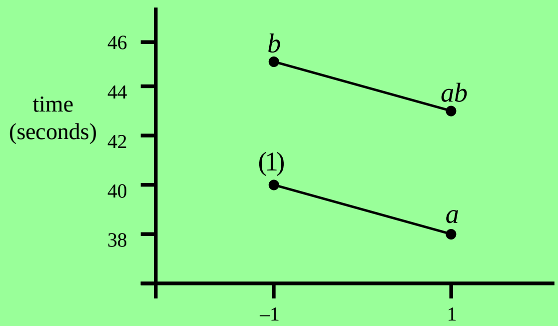
$$= \frac{38+43}{2} - \frac{40+45}{2} = 40.5 - 42.5 = \boxed{-2}$$

effect of B
(from –1 to 1)
(Jim to Rachel)

$$= \frac{b+ab}{2} - \frac{(1)+a}{2}$$

$$= \frac{45+43}{2} - \frac{40+38}{2} = 44 - 39 = \boxed{5}$$

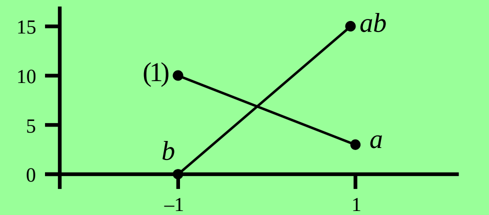
Interaction Plot



Interaction: when factors interact with each other in unexpected ways

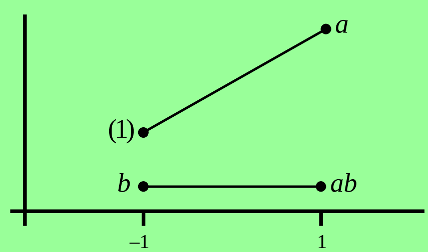
	$X_1 = -1$	$X_1 = 1$
$X_2 = 1$	b = 0	ab = 15
$X_2 = -1$	(1) = 10	a = 3

why not 25??



Any interaction plot that isn't "parallel" may indicate significant interaction.

another example:



Model

"intercept"

coefficients

$\varepsilon \sim N(0,\sigma)$
error

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

*i*th
observation

$\boxed{-1 \text{ or } 1}$

$$\widehat{\beta}_0 = \text{overall mean}$$

$$\widehat{\beta}_1 = \frac{\text{estimated effect for Factor A}}{2} = \frac{-(1)+a-b+ab}{4}$$

$$\widehat{\beta}_2 = \frac{\text{estimated effect for Factor B}}{2} = \frac{-(1)-a+b+ab}{4}$$

$$\widehat{\beta}_{12} = \frac{(1)-a-b+ab}{4}$$

Remote Control Cars: new experiment

response variable Y = time to complete course (seconds)

Jim & Rachel discard the octane factor and design a new experiment – this time the model will include interaction.

Factor A: Controller Design	Factor B: Operator
Levels: old or new	Levels: Jim or Rachel
"low" $x_1 = -1$, "high" $x_1 = 1$	"low" $x_2 = -1$, "high" $x_2 = 1$

<u>treatment combinations</u>:

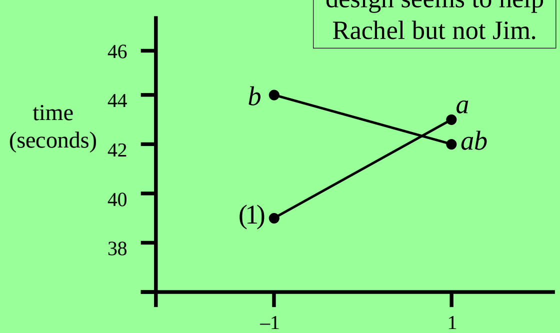
X_1 old or new	X_2 Jim or Rachel	X_1X_2	description	average
–1	–1	1	old/Jim	(1) = 39
1	–1	–1	new/Jim	a = 43
–1	1	–1	old/Rachel	b = 44
1	1	1	new/Rachel	ab = 42

run in random order

		Factor A: $X_1 = -1$ old	Factor A: $X_1 = 1$ new	
Factor B	$X_2 = 1$ Rachel	b = 44	ab = 42	43
Factor B	$X_2 = -1$ Jim	(1) = 39	a = 43	41
		41.5	42.5	42

<u>Interaction Plot</u>

The new controller design seems to help Rachel but not Jim.



(1) = 39	a = 43	b = 44	ab = 42

$$\hat{\beta}_0 = \text{overall mean} = 42$$

$$\hat{\beta}_1 = \frac{-(1)+a-b+ab}{4} = \frac{-39+43-44+42}{4} = \boxed{0.5}$$

$$\hat{\beta}_2 = \frac{-(1)-a+b+ab}{4} = \frac{-39-43+44+42}{4} = \boxed{1}$$

$$\hat{\beta}_{12} = \frac{(1)-a-b+ab}{4} = \frac{39-43-44+42}{4} = \boxed{-1.5}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$$

$$\hat{Y} = 42 + 0.5x_1 + x_2 - 1.5x_1x_2$$

<u>check</u>:

X_1	X_2	$\hat{Y} = 42 + 0.5x_1 + x_2 - 1.5x_1x_2$
–1	–1	$\hat{Y} = 42 + 0.5(-1) + (-1) - 1.5(-1)(-1) = 39$
1	–1	$\hat{Y} = 42 + 0.5(1) + (-1) - 1.5(1)(-1) = 43$
–1	1	$\hat{Y} = 42 + 0.5(-1) + (1) - 1.5(-1)(1) = 44$
1	1	$\hat{Y} = 42 + 0.5(1) + (1) - 1.5(1)(1) = 42$

MA 3710: World of Formula
Md Mutasim Billah

Descriptive Statistics:

$$\overline{X} = \frac{\sum x}{n}$$

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

$$s^2 = \frac{\sum\left(x - \overline{X}\right)^2}{n-1} = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

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

Locator numbers:

Q_1 if n odd	Q_1 if n even	Q_2	Q_3 if n odd	Q_3 if n even
$\frac{n+3}{4}$	$\frac{n+2}{4}$	$\frac{n+1}{2}$	$\frac{3n+1}{4}$	$\frac{3n+2}{4}$

IQR = $Q_3 - Q_1$

Fences:

LIF = $Q_1 - 1.5(\text{IQR})$
UIF = $Q_3 + 1.5(\text{IQR})$

LOF = $Q_1 - 3(\text{IQR})$
UOF = $Q_3 + 3(\text{IQR})$

Probability:

$P(A) = 1 - P(\overline{A})$

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

A and B are mutually exclusive iff $P(A \cap B) = 0$

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

Two events A and B are independent iff
$P(A|B) = P(A)$ iff
$P(B|A) = P(B)$ iff
$P(A \cap B) = P(A) \cdot P(B)$

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(\overline{A}) \cdot P(B|\overline{A})}$$

$$nPr = \frac{n!}{(n-r)!}$$

$$nCr = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

Probability distribution:

	discrete	continuous
$\boldsymbol{\mu = E(X)}$	$\sum x \cdot p(x)$	$\int_{-\infty}^{\infty} x \cdot f(x)\,dx$
$\boldsymbol{\sigma^2 = Var(X)}$	$\sum (x-\mu)^2 \cdot p(x)$ $E(X^2) - \mu^2$	$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx$ $E(X^2) - \mu^2$
$\boldsymbol{\sigma = SD(X)}$	$\sqrt{\sigma^2}$	$\sqrt{\sigma^2}$
$\boldsymbol{M(t) = E(e^{tX})}$	$\sum e^{tX} \cdot p(x)$	$\int_{-\infty}^{\infty} e^{tX} \cdot f(x)\,dx$

Famous *Discrete* Random Variables:

name	pmf	mean	variance	mgf
Binomial $X \sim BIN(n,p)$	$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0,1,2,\dots n \\ 0 & \text{otherwise} \end{cases}$	np	$np(1-p)$	$[pe^t + (1-p)]^n$
Geometric $X \sim GEO(p)$	$p(x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1,2,\dots \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Negative Binomial $X \sim NEGBIN(p,r)$	$p(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{if } x = r, r+1, r+2, \dots \\ 0 & \text{otherwise} \end{cases}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$
Poisson $X \sim POI(\lambda)$	$p(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x = 0,1,2,\dots \\ 0 & \text{otherwise} \end{cases}$	λ	λ	$e^{\lambda(e^t - 1)}$
Hypergeometric $X \sim HYPGEO(N,n,r)$	$p(x) = \begin{cases} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} & \text{if } x = 0,1,2,\dots n \text{ and } n \le r \\ & \text{or } x = 0,1,2,\dots r \text{ and } n > r \\ 0 & \text{otherwise} \end{cases}$	$\frac{nr}{N}$	$\frac{n\,r\,(N-r)\,(N-n)}{N^2\,(N-1)}$	Does not exist in closed form.

Famous *Continuous* Random Variables:

name	pdf	mean	variance	mgf
Uniform $X \sim UNIF(\theta_1, \theta_2)$	$f(x) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 < x < \theta_2 \\ 0 & \text{otherwise} \end{cases}$	$\frac{\theta_1 + \theta_2}{2}$	$\frac{(\theta_2 - \theta_1)^2}{12}$	$\frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$
Normal $X \sim N(\mu, \sigma)$	$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & \text{if } -\infty < x < \infty \end{cases}$	μ	σ^2	$e^{\left(\mu t + \frac{t^2\sigma^2}{2}\right)}$
Exponential $X \sim EXP(\lambda)$ $X \sim GAMMA\left(\alpha = 1, \beta = \frac{1}{\lambda}\right)$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\lambda}$	$\left(\frac{1}{\lambda}\right)^2$	$\left(1 - \left(\frac{1}{\lambda}\right)t\right)^{-1}$
Gamma $X \sim GAMMA(\alpha, \beta)$	$f(x) = \begin{cases} \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\,\beta^\alpha} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta t)^{-\alpha}$
Chi-Square $X \sim \chi^2(v)$ $X \sim GAMMA\left(\alpha = \frac{v}{2}, \beta = 2\right)$	$f(x) = \begin{cases} \frac{x^{(v/2 - 1)} e^{-\frac{x}{2}}}{\Gamma\left(\frac{v}{2}\right) 2^{(v/2)}} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$	v	$2v$	$(1-2t)^{-\frac{v}{2}}$
Beta $X \sim BETA(\alpha, \beta)$	$f(x) = \begin{cases} \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\right] x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}$	Does not exist in closed form.

The Normal Distribution:

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

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Confidence Interval:

unknown parameter	confidence interval	requirements	sample size
μ	$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	σ known normal population or $n \geq 30$	$n = \left(\frac{Z_{\frac{\alpha}{2}}\ \sigma}{B} \right)^2$
μ	$\overline{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	normal population or $n \geq 30$	
p	$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$n \geq 30$	$n = \hat{p}(1-\hat{p}) \left(\frac{Z_{\frac{\alpha}{2}}}{B} \right)^2$

Hypothesis Testing:

unknown parameter	test statistic	requirements
μ	$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$	σ known normal population or $n \geq 30$
μ	$T = \frac{\overline{X} - \mu}{s / \sqrt{n}}$	normal population or $n \geq 30$
p	$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$n \geq 30$

Hypothesis Testing (continue..):

unknown quantity	test statistic	requirements
μ_d	$T = \dfrac{\overline{X}_d - \mu_d}{s_d / \sqrt{n_d}}$ d.o.f. $= n_d - 1$	population of differences is normal or large sample
$\mu_1 - \mu_2$	$T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ d.o.f. $= n_1 + n_2 - 2$	independent samples normal populations or large samples $\sigma_1^2 = \sigma_2^2$
$p_1 - p_2$	$Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$	independent samples large samples

Control Chart:

Chart	LCL	UCL
$\overline{X}$	$\overline{\overline{X}} - 3\sqrt{\dfrac{\overline{s^2}}{n}}$	$\overline{\overline{X}} + 3\sqrt{\dfrac{\overline{s^2}}{n}}$
s^2	$\dfrac{\left(\chi^2_{n-1,0.99865}\right)\overline{s^2}}{n-1}$	$\dfrac{\left(\chi^2_{n-1,0.001355}\right)\overline{s^2}}{n-1}$
X	$\overline{X} - 2.66 \cdot \overline{MR}$	$\overline{X} + 2.66 \cdot \overline{MR}$
np	$n\overline{p} - 3\sqrt{n\overline{p}(1-\overline{p})}$	$n\overline{p} + 3\sqrt{n\overline{p}(1-\overline{p})}$
c	$\overline{c} - 3\sqrt{\overline{c}}$	$\overline{c} + 3\sqrt{\overline{c}}$

<table>
<tr><th>chart</th><th>assumptions</th><th>Phase 2 alarm?</th><th>Phase 1 alarm?</th></tr>
<tr><td>$\overline{X}$</td><td rowspan="2">*Process variance is constant
*in control during Phase 1
*sample means are normally distributed</td><td>Evidence that process is out of control.</td><td rowspan="5">Evidence that assumption(s) are violated – chart is invalid. Start over.</td></tr>
<tr><td>s^2</td><td>Evidence that assumption(s) are violated – chart is invalid. Start over.</td></tr>
<tr><td>X</td><td>*in control during Phase 1
*data is normally distributed</td><td rowspan="3">Evidence that process is out of control.</td></tr>
<tr><td>np</td><td>*in control during Phase 1
*data is Binomial</td></tr>
<tr><td>c</td><td>*in control during Phase 1
*data is Poisson</td></tr>
</table>

Regression Analysis:

$SSxx = \sum x^2 - \frac{(\sum x)^2}{n}$	$r = \frac{SSxy}{\sqrt{SSxx}\sqrt{SSyy}}$
$SSyy = \sum y^2 - \frac{(\sum y)^2}{n}$	$\hat{\beta}_1 = \frac{r \cdot s_y}{s_x} = \frac{SSxy}{SSxx}$
$SSxy = \sum xy - \frac{(\sum x)(\sum y)}{n}$	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$
	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$SS_{total} = Syy$	$r^2 = \frac{SSR}{SS_{total}}$
$SSR = \hat{\beta}_1 (Sxy)$	$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{SSE}{(n-2)Sxx}}}$
$SSE = SS_{total} - SSR$	

One-way ANOVA:

$$MSA = \frac{SSA}{k-1} = \frac{\sum_{i=1}^{k} n_i\left(\overline{X}_i - \overline{X}\right)^2}{k-1}$$

$$MSE = \frac{SSE}{n-k} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k}$$

$$F = \frac{MSA}{MSE} = \frac{\left(\frac{SSA}{k-1}\right)}{\left(\frac{SSE}{n-k}\right)}$$

numerator d.o.f. $= k - 1$
denominator d.o.f. $= n - k$

requirements:
The k samples are independent of each other.
The populations are normally distributed.
All the population variances are the same:
$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$

Source	d.o.f.	SS	MS	F
Treatment	$k-1$	SSA	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SS Total		

Fisher's LSD $= t_{n-k,\alpha/2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

Chi-square Test:

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

One variable test degrees of freedom:	$k - 1$
Two variable test degrees of freedom:	$(r-1)(c-1)$

Expected table for two variable test:

Outside boxes = same as observed table
inside boxes = $\frac{(\text{row total})(\text{column total})}{(\text{overall total})}$

Two-way ANOVA:

Source	d.o.f.	SS	MS	F
A (treatment)	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSE}$
B (block)	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F = \frac{MSB}{MSE}$
Error	$abn - a - b + 1$	SSE	$MSE = \frac{SSE}{abn-a-b+1}$	
Total	$abn - 1$	SS Total		

Fisher's LSD $= t_{abn-a-b+1,\alpha/2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{\frac{1}{n_\alpha} + \frac{1}{n_\beta}}$

$n_\alpha = n_\beta = nb$ if comparing row (treatment) means
$n_\alpha = n_\beta = na$ if comparing column (block) means

requirements:
The treatment combination samples are independent of each other.
The data is normally distributed.
All the treatment combination variances are the same.
There is no interaction between the two factors.

Engineering Statistics

Chapter 05: control charts

Video 12.01
introduction to control charts

Department of Mathematical Sciences
Philip Kendall, Senior Lecturer

Michigan Tech 1885

Control charts

Engineers are frequently required to monitor a process, especially in the context of manufacturing.

Regardless of how much care and expertise is devoted to designing and maintaining a process, eventually a certain amount of unwanted variability is inevitable because machines and their parts wear down, workers make mistakes, materials and other inputs to the process are imperfect and change over time, and so on.

A control chart enables an engineer to monitor a process. When the amount of variability becomes unacceptable, the engineer can interrupt the process and take appropriate action.

Dairy Example

Plastic milk jugs are labelled as containing 1 gallon of milk. Of course, not every milk jug contains *exactly* 1.0000000000 gallons.

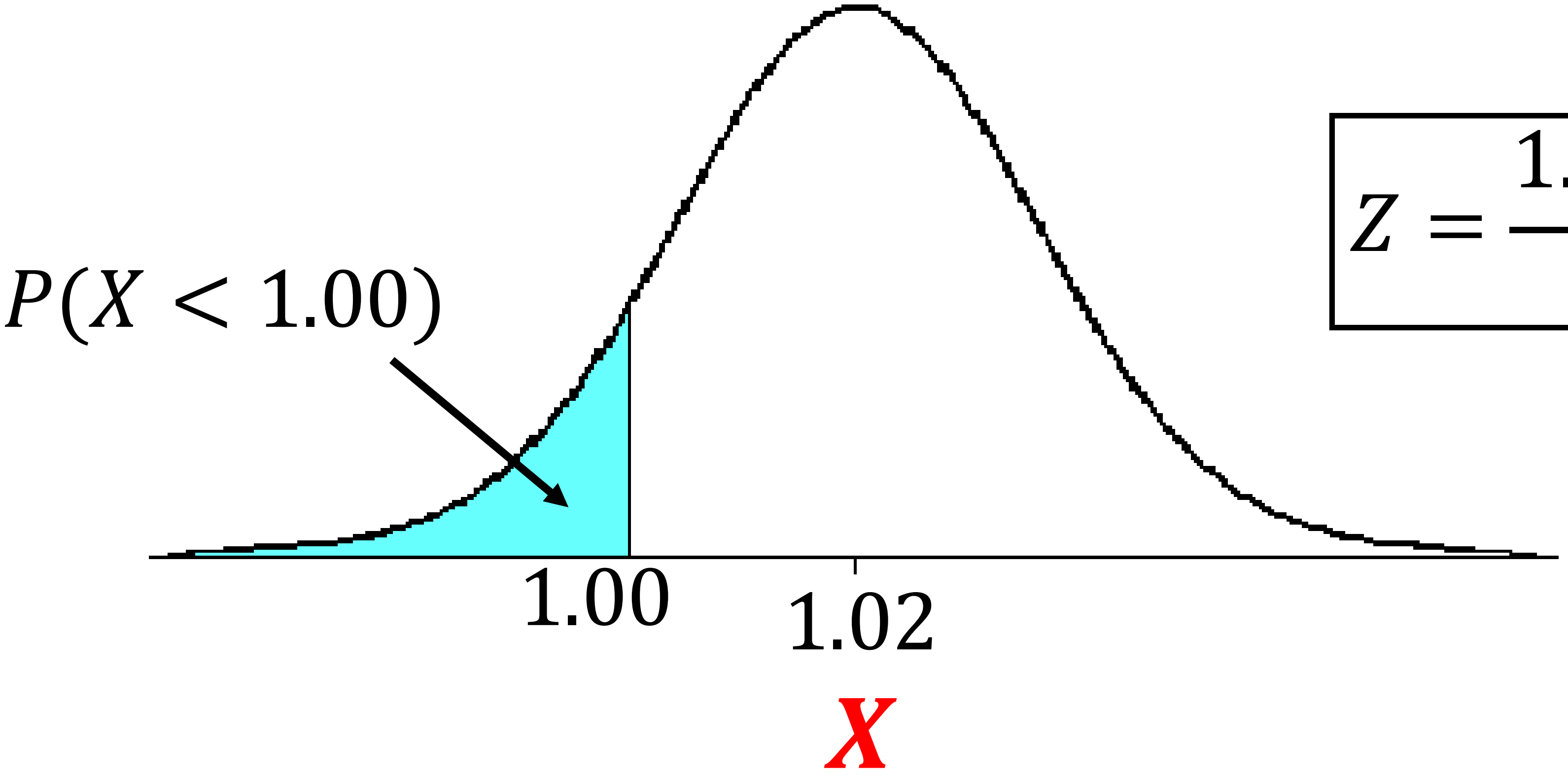
Milk jugs should not be underfilled because that would lead to customer dissatisfaction and potential legal issues. Milk jugs should not be overfilled because that would reduce profitability and possibly threaten the capacity of the milk jugs themselves.

Engineers have designed a process such that the milk jugs are filled with an amount of milk that is a normally distributed random variable:

$$X \sim N(\mu = 1.02, \sigma = .01)$$

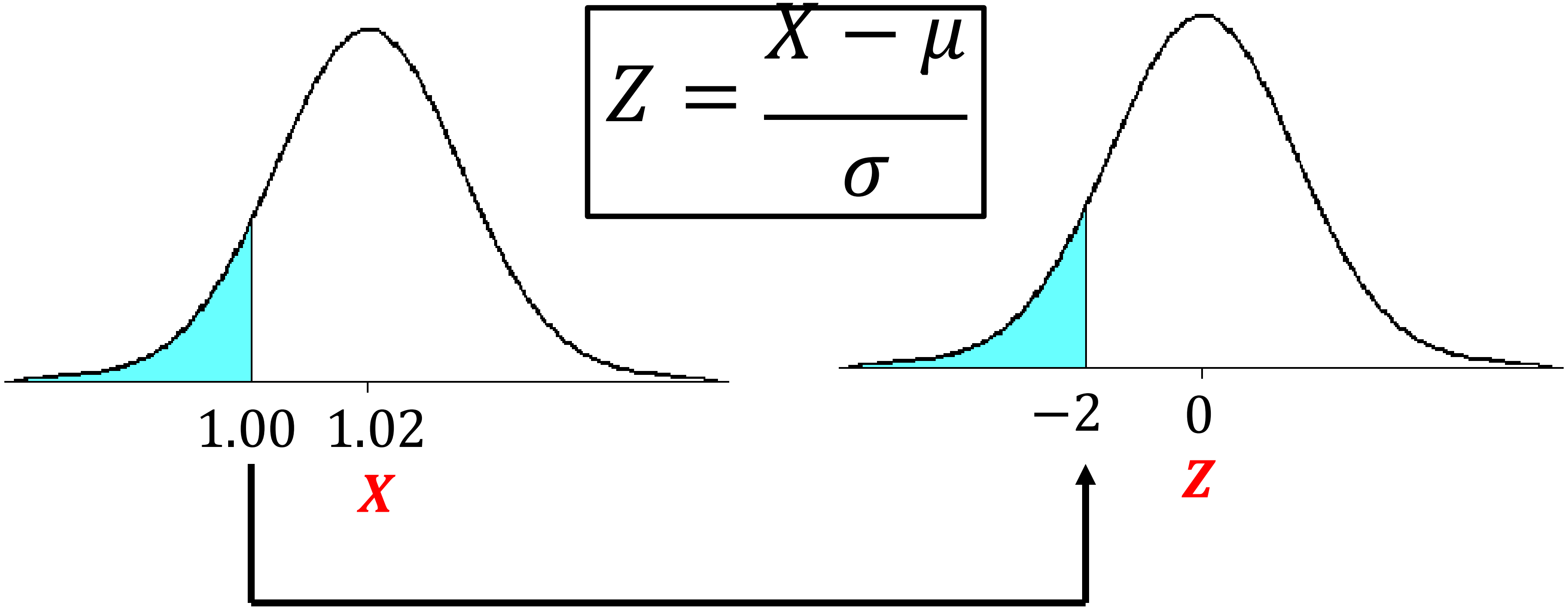
X = amount of milk dispensed into a milk jug

$X \sim N(\mu = 1.02 \text{ gallons}, \sigma = .01)$



$$\boxed{Z = \frac{1.00 - 1.02}{.01} = -2}$$

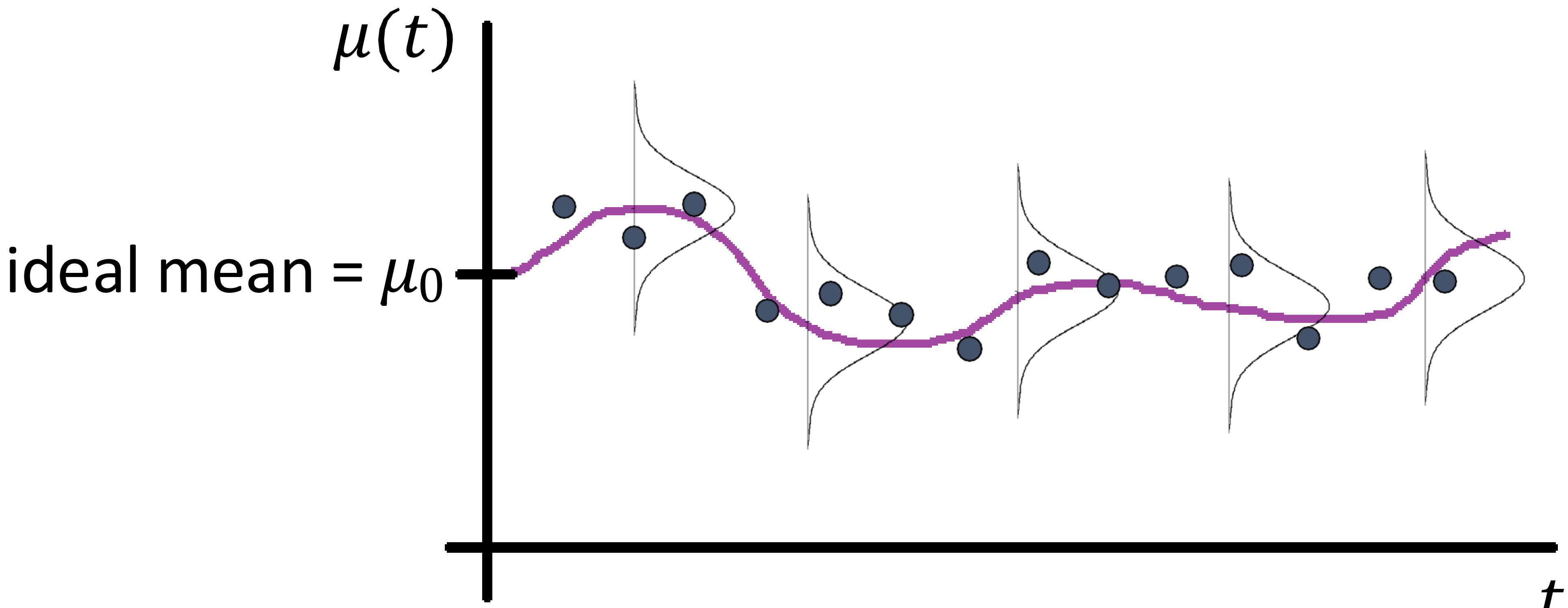
$$\boxed{Z = \frac{X - \mu}{\sigma}}$$



$$P(X < 1.00) = P\left(Z < \frac{1.00 - 1.02}{.01}\right) = P(Z < -2) = .0228 \approx \frac{1}{44}$$

It is accepted that in the long run about 1 in 44 milk jugs will be underfilled.

The mean of a process can change over time.



$$X_i = \mu(t) + \epsilon_i$$

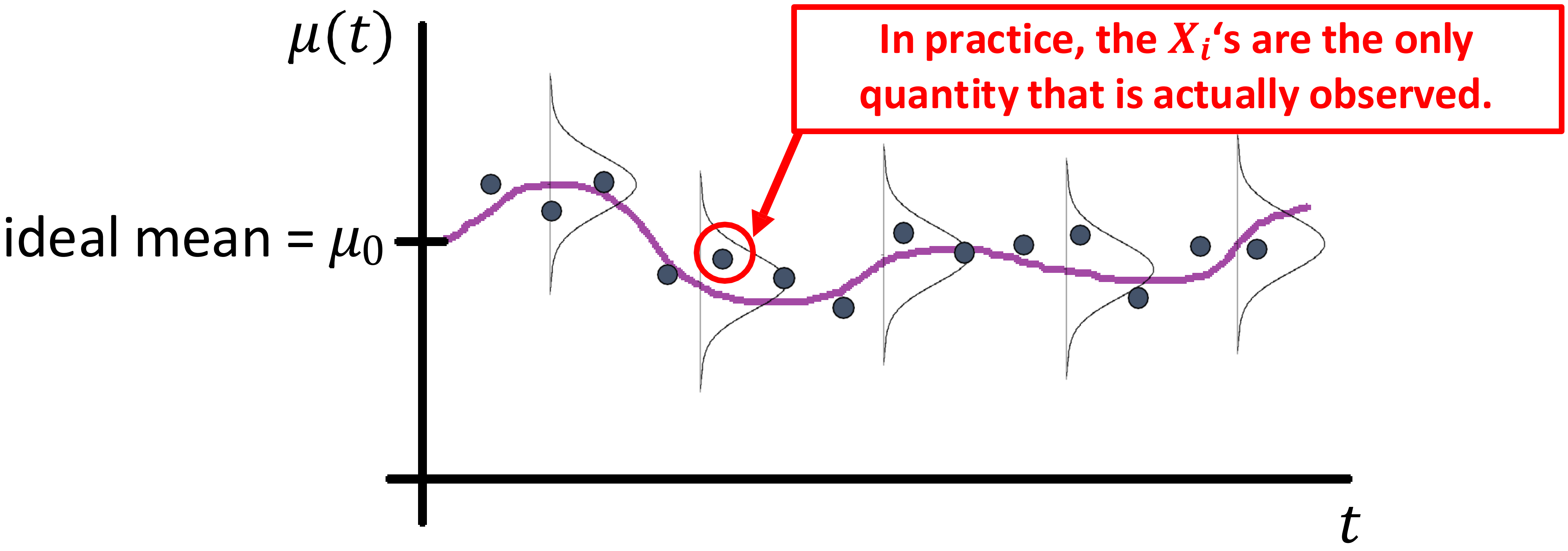
$\epsilon_i \sim N(0, \sigma)$

***i*th observation**

mean at time *t*

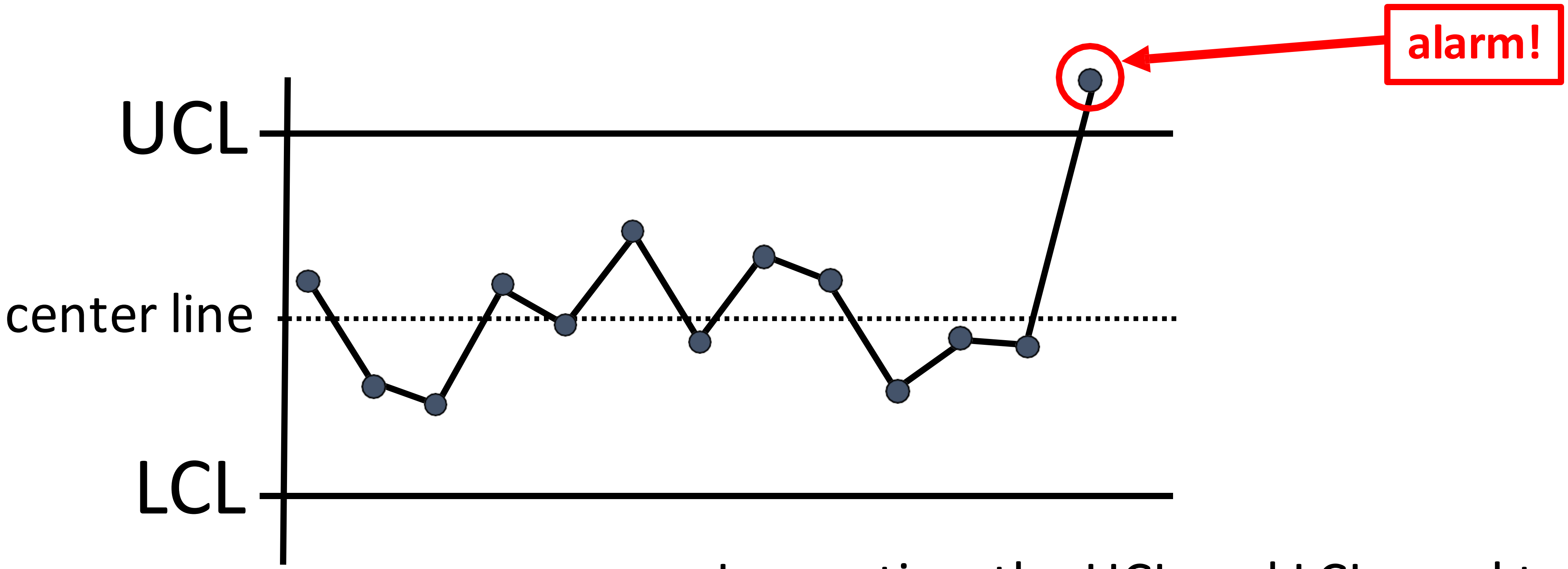
error

assumption: errors are normal with constant variability



The value of $\mu(t)$ will not be known for all t. In fact, $\mu(t)$ is typically never known with complete certainty and can only be estimated periodically, usually at fixed time intervals.

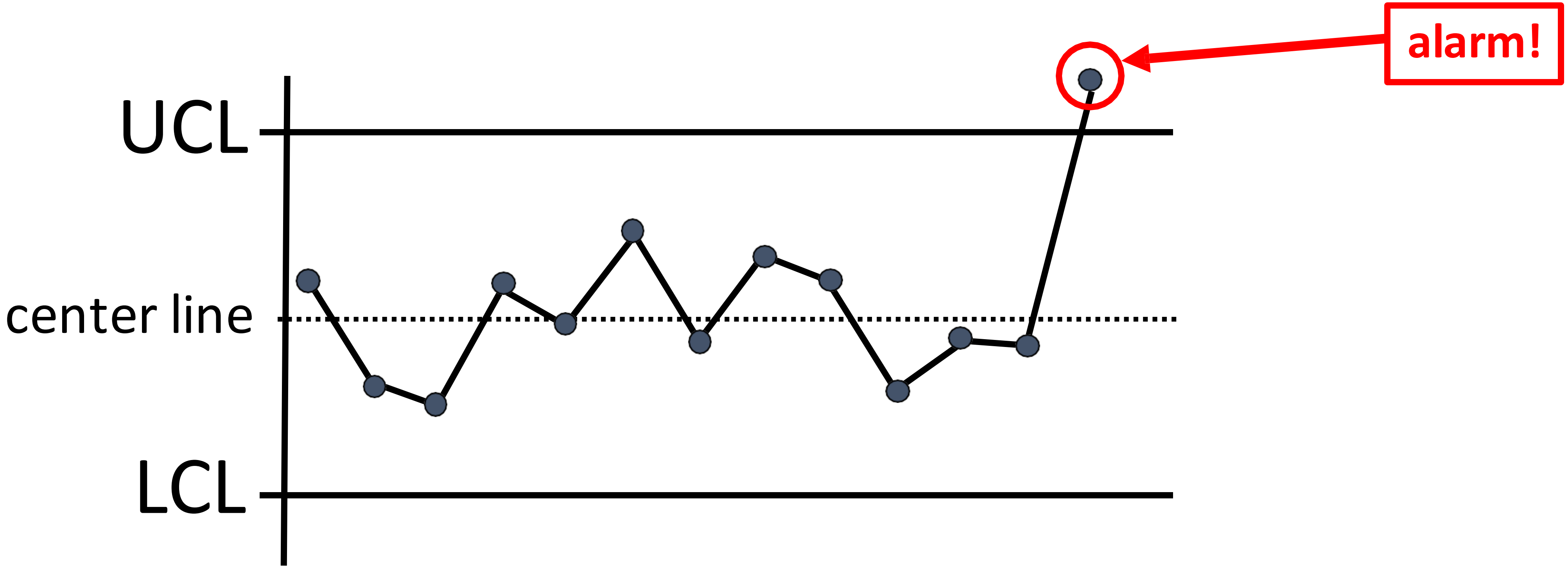
Idea: Take random samples from the process at fixed time intervals and plot $\overline{X}$ (or whatever statistic is appropriate).



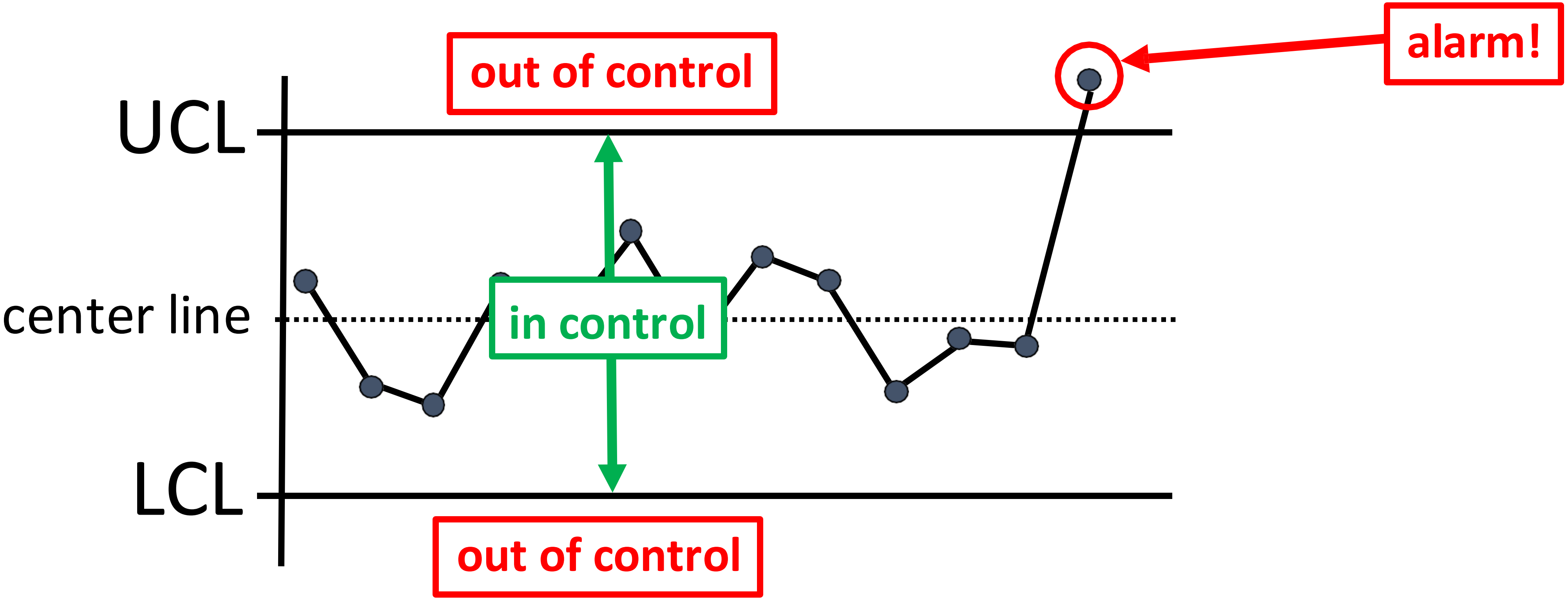
In practice, the UCL and LCL need to estimated with sample data collected during a Phase 1 "base period".

UCL = upper control limit

LCL = lower control limit



The center line can be a theoretical mean determined by engineers, but often the center line also needs to be estimated with sample data collected during the Phase 1 “base period”.



Think of each data point on the control chart as a hypothesis test, where the null hypothesis claims the process is in control. An observation outside the control limits (an "alarm") is a rejection of the null hypothesis and a conclusion that the process is now out of control.

Back to the Dairy Example

Engineers have designed a process such that the milk jugs are filled with an amount of milk that is a normally distributed random variable:

$$X \sim N(\mu = 1.02, \sigma = .01)$$

They have decided to randomly select n = 4 milk jugs per day and plot $\overline{X}$ on a control chart.

They have decided to randomly select n = 4 milk jugs per day and plot $\overline{X}$ on a control chart.

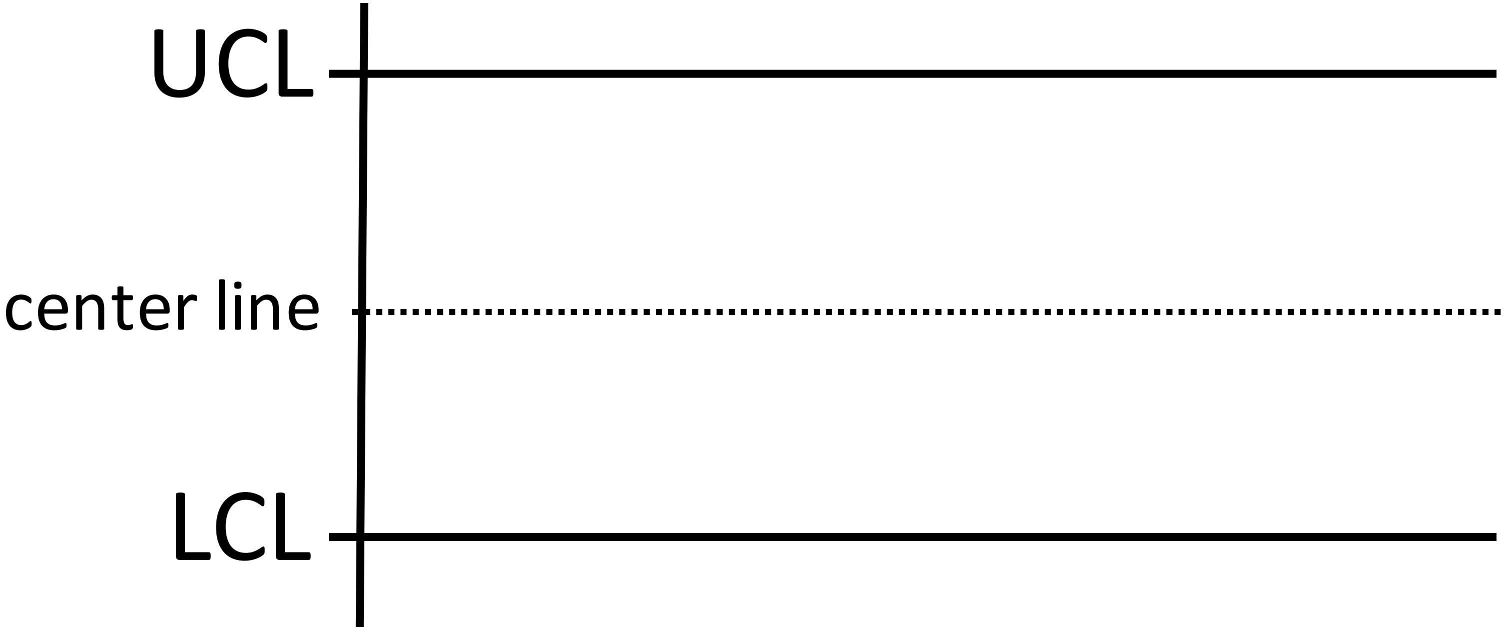
This is a theoretical and unrealistic example because the engineers have many knowns and few unknowns:

<u>known</u>:

$$\mu(t) = 1.02 \text{ at } t = 0$$
$$\sigma = .01$$
data is normally distributed
$$X \sim N(\mu(t), \sigma = .01)$$
$$\overline{X} \sim N\left(\mu(t), \frac{\sigma}{\sqrt{n}} = \frac{.01}{\sqrt{4}}\right)$$

<u>unknown</u>:

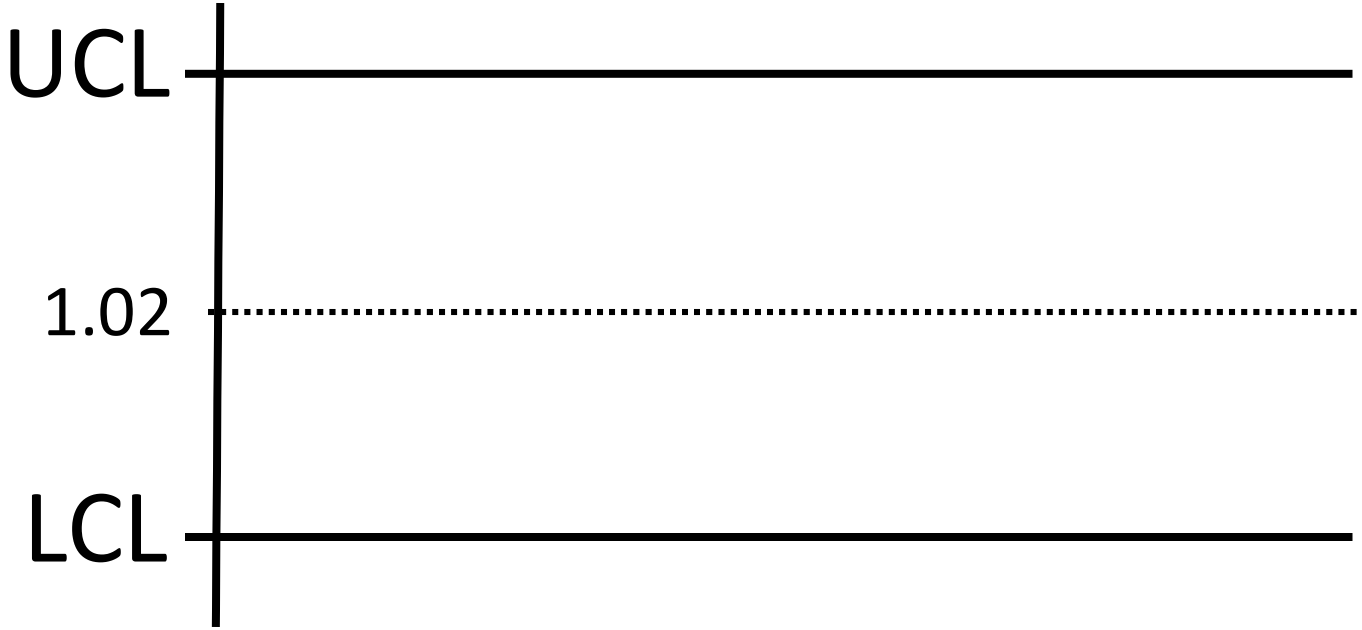
$$\mu(t) \text{ when } t > 0$$



center line = μ_0 = 1.02 gallons

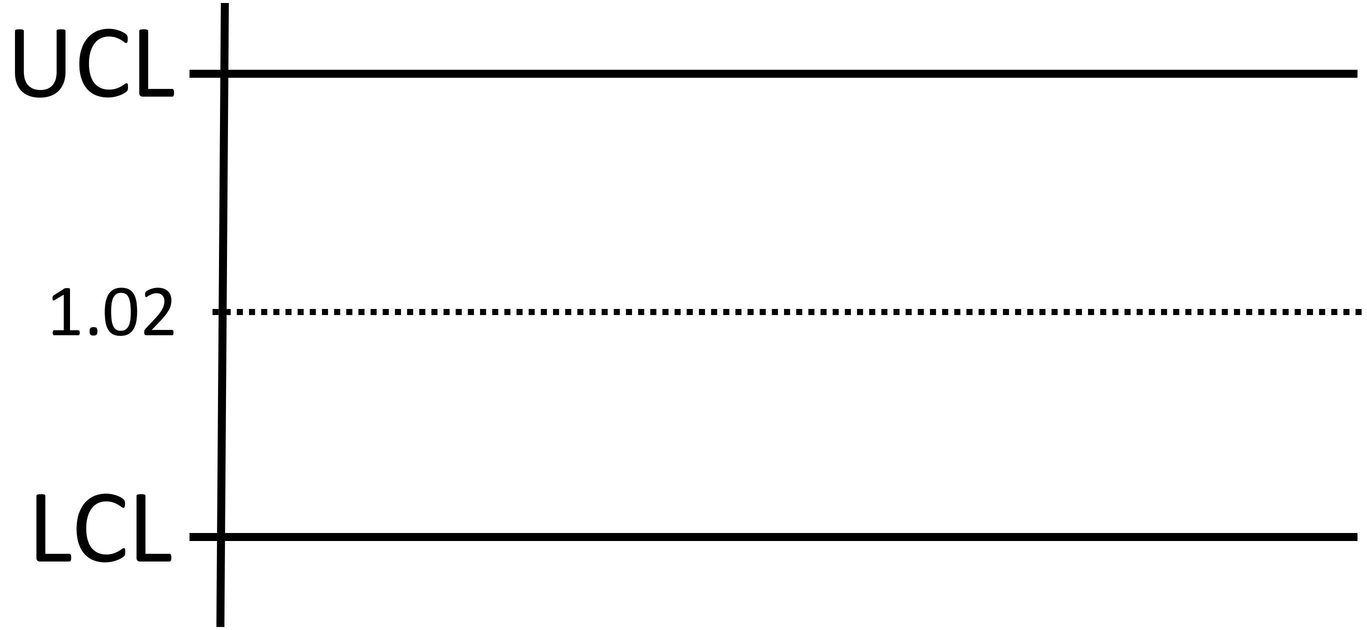
Historically, engineers choose to place the upper and lower control limits 3 standard deviations above and below the center line.

$$\{ \text{LCL}, \text{UCL} \} = \text{center line} \pm 3 \cdot \text{SD(observed statistic)}$$



{ LCL, UCL } = center line $\pm$ 3·SD(observed statistic)

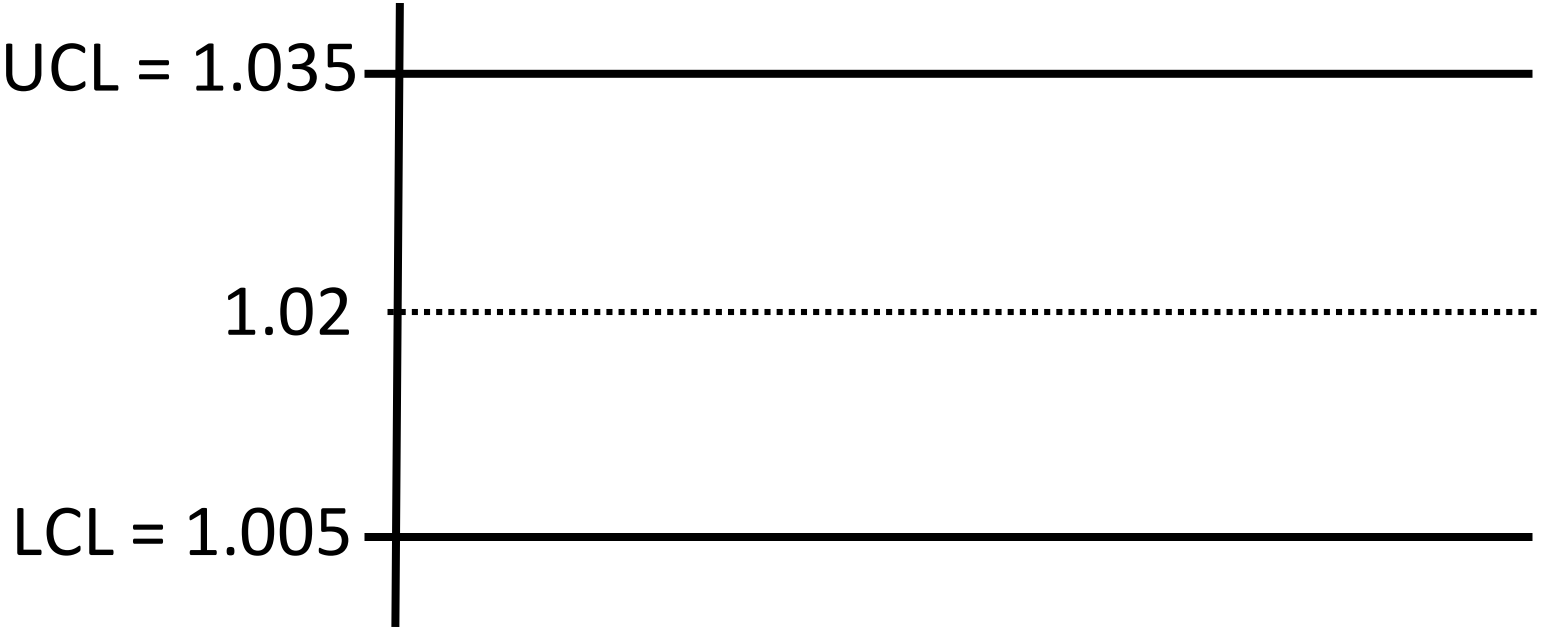
For this control chart, the observed statistic is $\overline{X}$ and has a standard deviation of $\frac{\sigma}{\sqrt{n}} = \frac{.01}{\sqrt{4}}$



$\{ \text{LCL}, \text{UCL} \} = \text{center line} \pm 3 \cdot \text{SD(statistic that is observed)}$

$$1.02 \pm 3\left(\frac{.01}{\sqrt{4}}\right) \rightarrow 1.02 \pm 0.015 = \boxed{\{ 1.005, 1.035 \}}$$

This is <u>not</u> a confidence interval! This is a set of two numbers.



Plan: randomly select n = 4 milk jugs per day and plot $\overline{X}$.

Each point on the control chart is a hypothesis test:

H_0: process mean is still 1.02 gallons

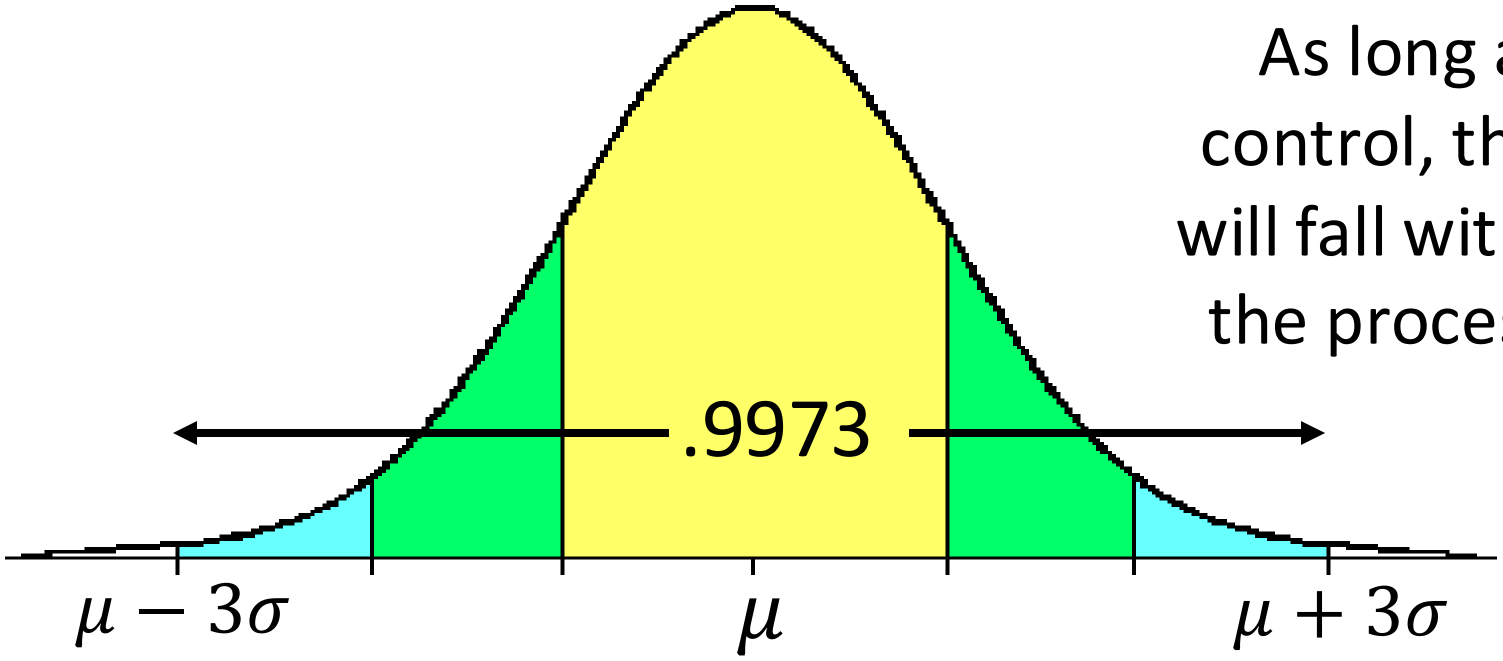
H_a: process mean is no longer 1.02 gallons (the process is now “out of control”)

Reject H_0 iff an $\overline{X}$ is above UCL or below LCL.

Time out:

$\{ \text{LCL}, \text{UCL} \} = \text{center line} \pm 3 \cdot \text{SD}(\text{statistic that is observed})$

Why <u>three</u> standard deviations?

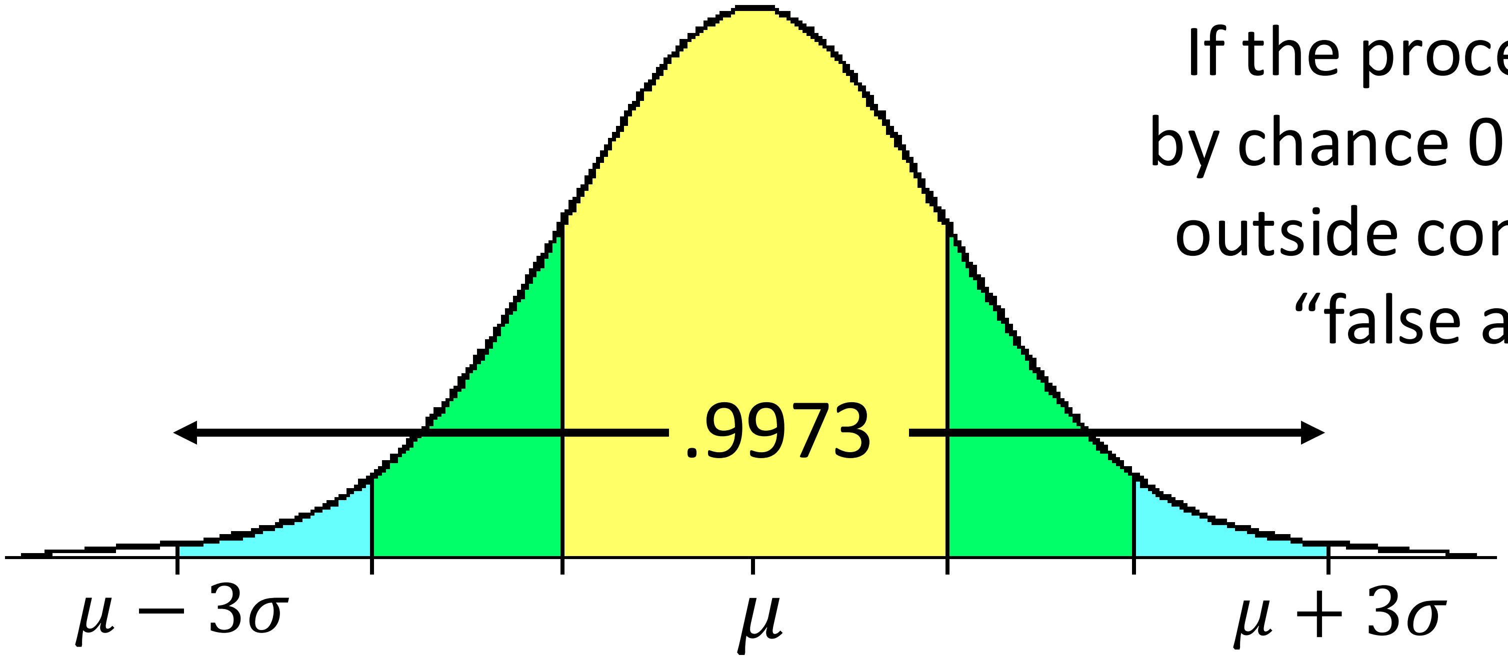


As long as the process remains in control, then 99.73% of observations will fall within 3 standard deviations of the process mean and not trigger an alarm.

Time out:

$$\{\, LCL, UCL \,\} = \text{center line} \pm 3 \cdot SD(\text{statistic that is observed})$$

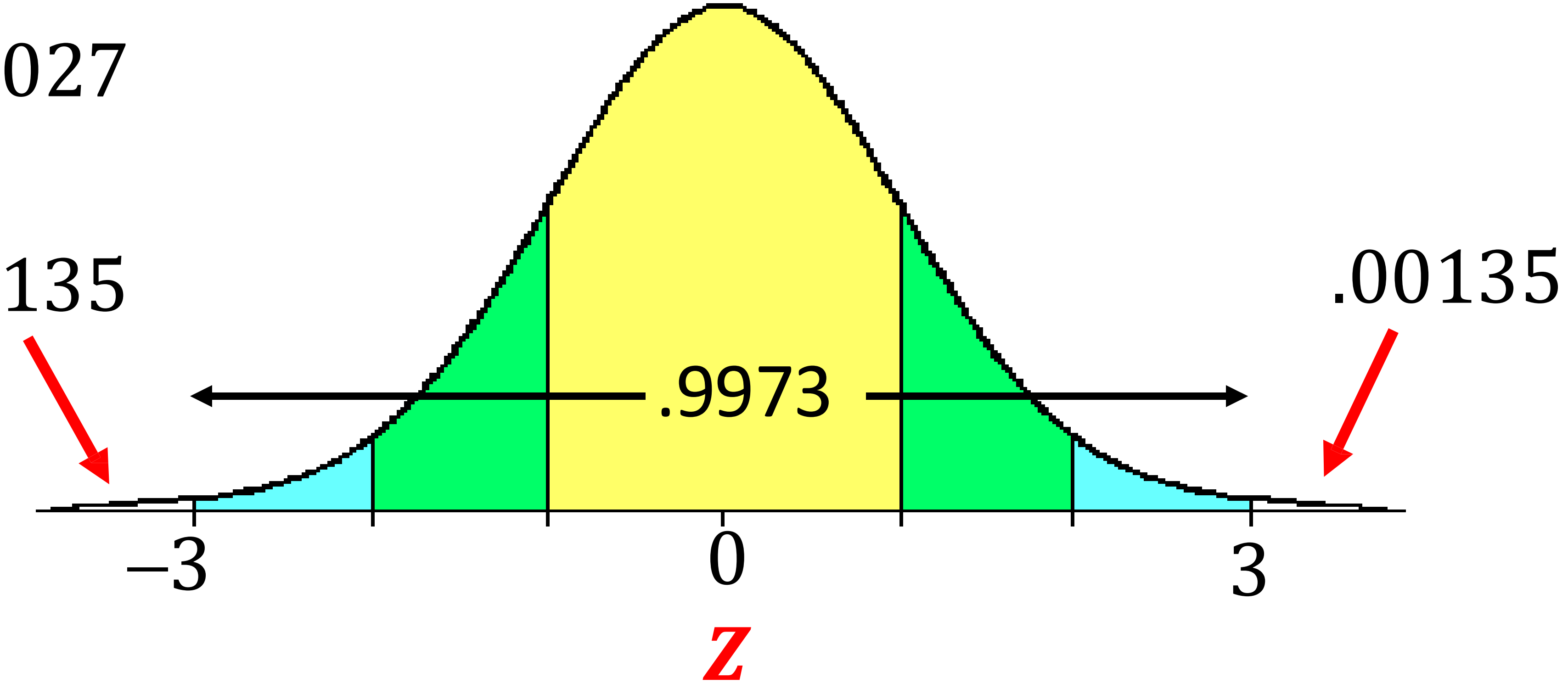
Why <u>three</u> standard deviations?



If the process remains in control, then by chance 0.27% of observations will fall outside control limits. This would be a "false alarm" or a Type 1 error.

$\alpha = .0027$

$\frac{\alpha}{2} = .00135$

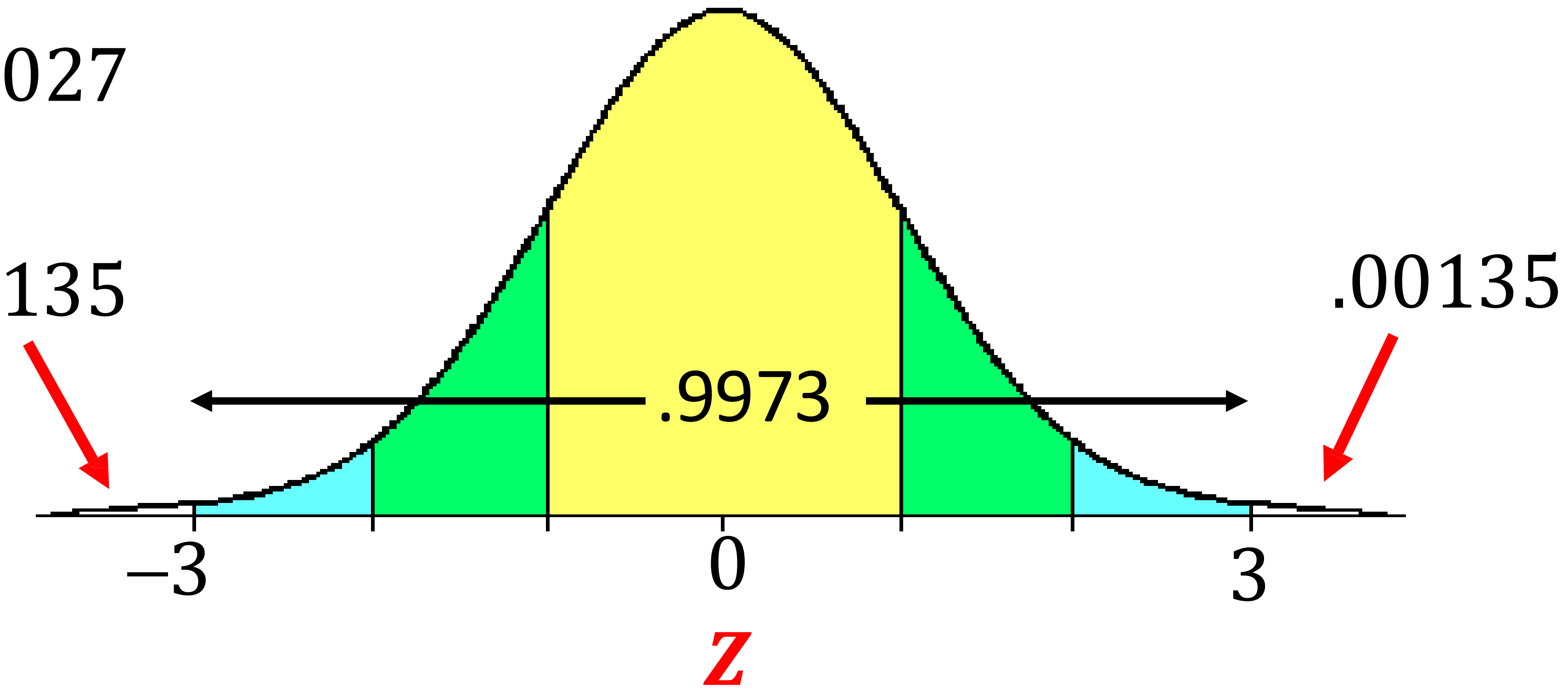


$$.0027 \approx \frac{1}{370}$$

It is important to remember that a process that is perfectly in control will inevitably false alarm, and it will occur about once in every 370 samples.

$\alpha = .0027$

$\frac{\alpha}{2} = .00135$



Why didn't we choose an α of .05?

Because it would result in too many false alarms. Since we will be conducting many hypothesis tests instead of just one, we need to keep α low to avoid making a lot of Type 1 errors.

Chart A: Process in control for all 20 days

$$\mu(t) = 1.02 \text{ for } 0 \le t \le 20$$

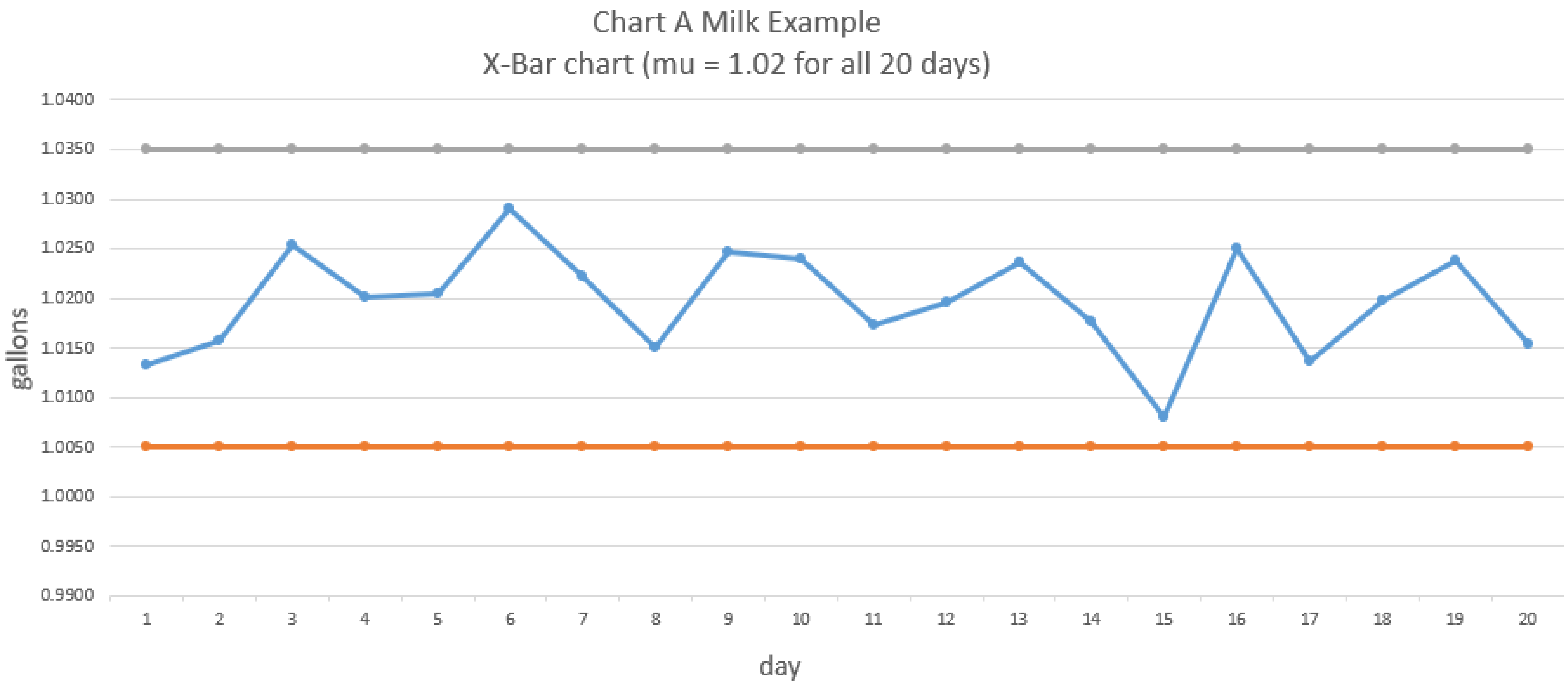
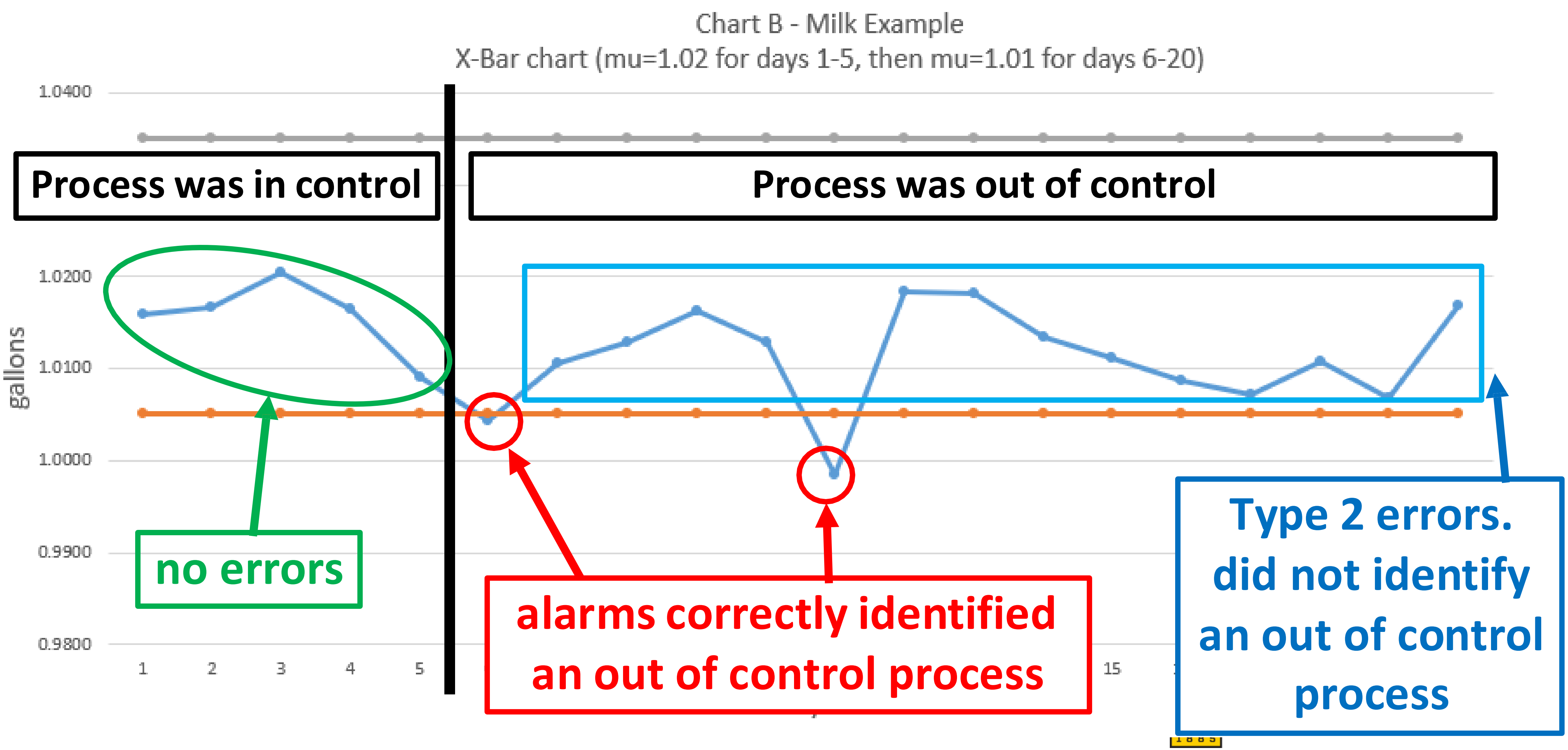


Chart B: Process out of control after Day 5

$$\mu(t) = 1.02 \text{ for } 1 \le t \le 5\text{, then } \mu(t) = 1.01 \text{ for } 6 \le t \le 20$$



Realistic situation:

We usually don't know $\mu(t)$ (even at $t = 0$) or σ.

Phase 1 (m samples)

Estimate any relevant parameters and compute the control limits with the estimates.
(We must assume the process is in control during Phase 1 or the estimates will not be reliable.)

Phase 2

Continue to monitor the process.

Engineering Statistics

Chapter 5: control charts

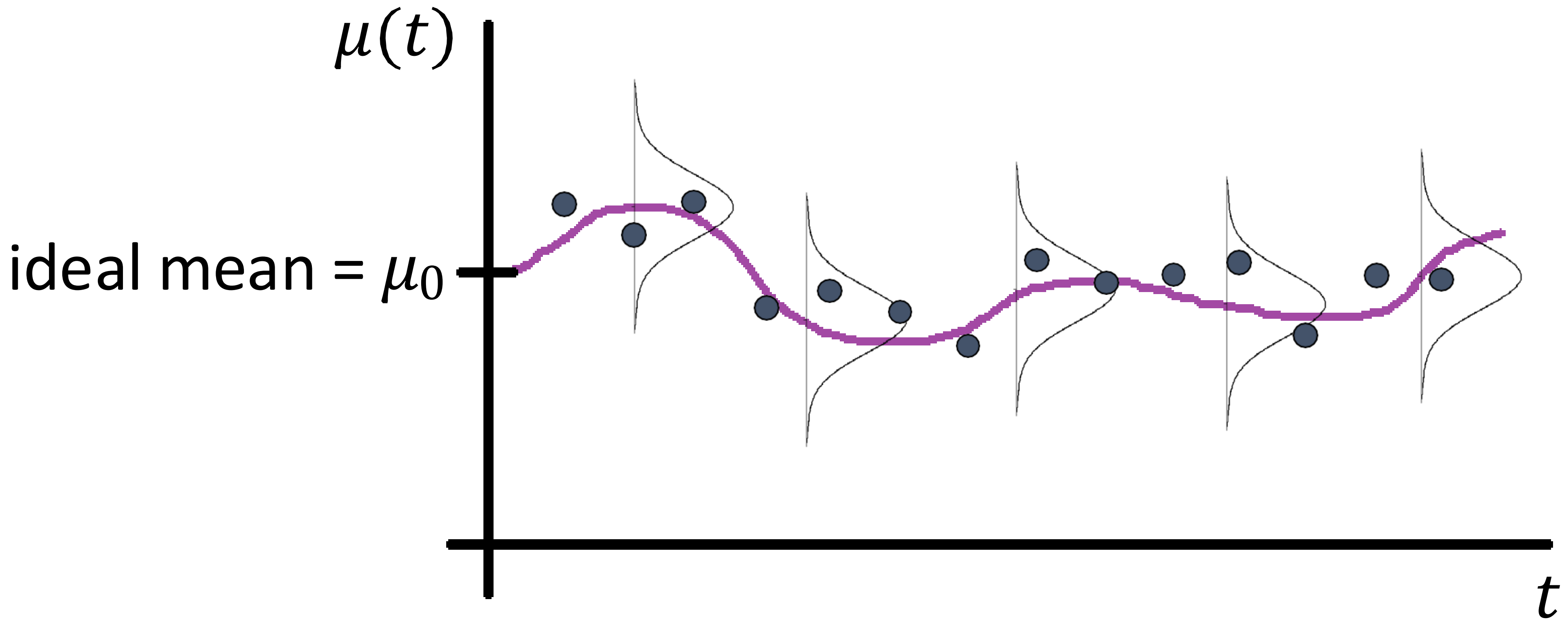
Video 12.02

$\bar{X}$ and s^2 charts

Department of Mathematical Sciences

Philip Kendall, Senior Lecturer

Michigan Tech



$$\epsilon_i \sim N(0, \sigma)$$

$$X_i = \mu(t) + \epsilon_i$$

ith observation

mean at time t

error

assumption: errors are normal with constant variability

An $\overline{X}$ chart monitors <u>between-subgroup</u> variability and assumes that σ^2 (the <u>within-group</u> variability) is constant.

So if σ^2 changes over time, the $\overline{X}$ chart is useless and unreliable. Therefore σ^2 needs to be monitored with its own chart and <u>any</u> alarms on the s^2 chart will indicate a violation of our assumption of constant variance.

<u>Realistic situation</u>:
μ and σ^2 are unknown and will be estimated during a Phase 1 base period.

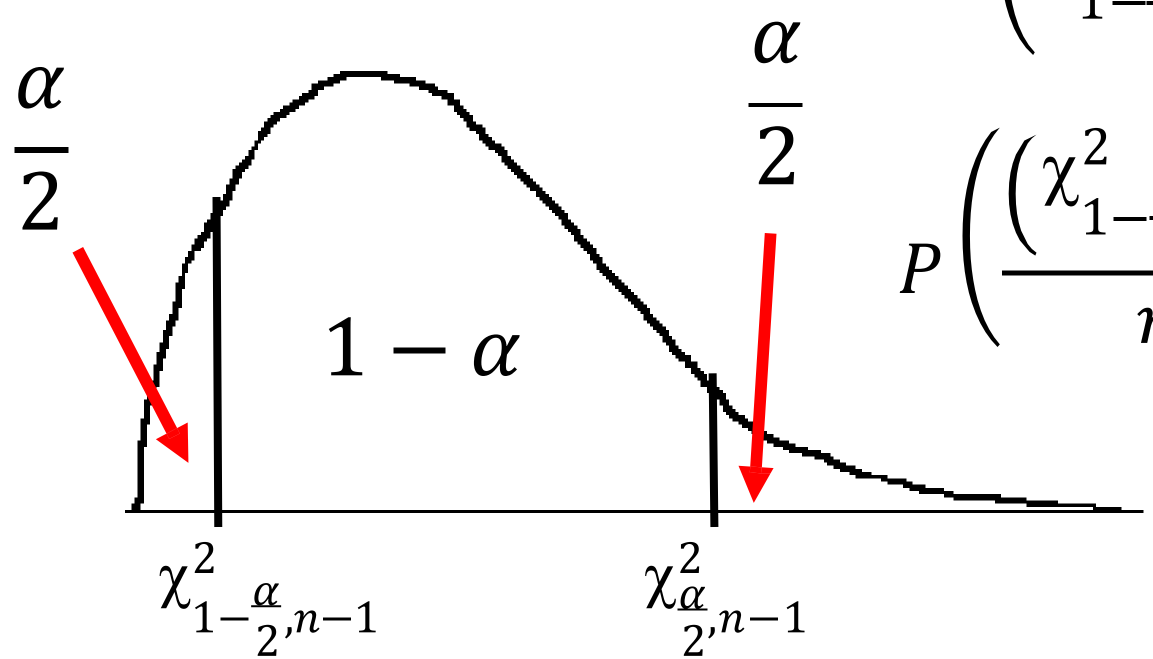
Useful fact:

If you are sampling from a normally distributed population, then

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

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

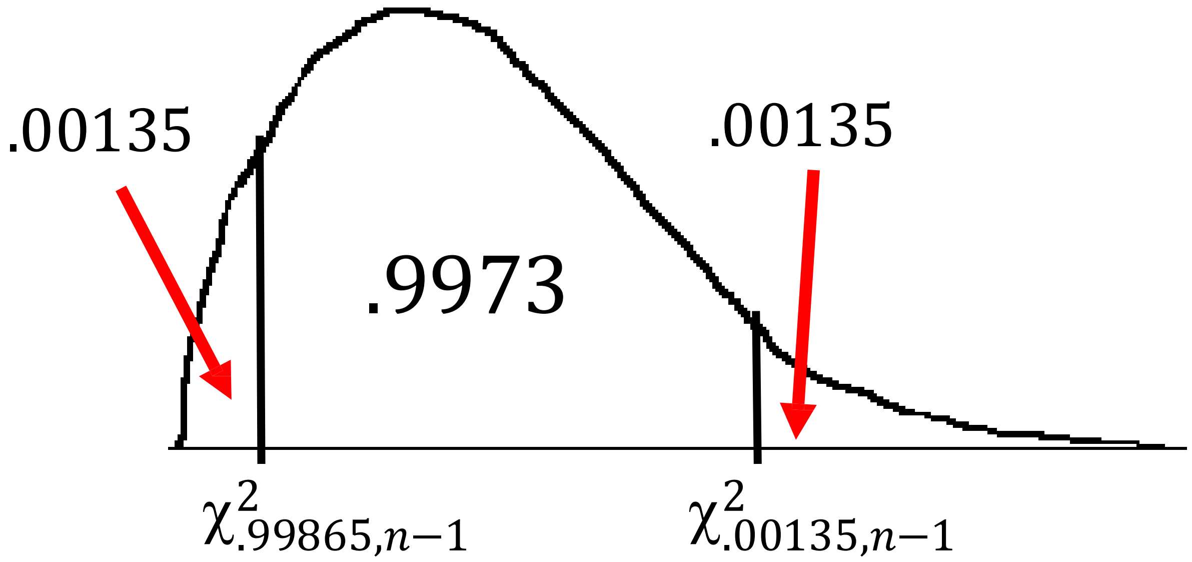
$$P\left(\frac{\left(\chi^2_{1-\frac{\alpha}{2},n-1}\right)\sigma^2}{n-1} < s^2 < \frac{\left(\chi^2_{\frac{\alpha}{2},n-1}\right)\sigma^2}{n-1}\right) = 1-\alpha$$



Previous video:

Historically, engineers like to use $\alpha = .0027$

$$P\left(\frac{\left(\chi^2_{.99865,n-1}\right)\sigma^2}{n-1} < s^2 < \frac{\left(\chi^2_{.00135,n-1}\right)\sigma^2}{n-1}\right) = .9973$$



$$P\left(\boxed{\frac{\left(\chi^2_{.99865,n-1}\right)\sigma^2}{n-1}} < s^2 < \boxed{\frac{\left(\chi^2_{.00135,n-1}\right)\sigma^2}{n-1}}\right) = .9973$$

σ^2 is unknown and will be estimated with the average sample variance during the Phase 1 base period.

s^2 chart control limits:

$$\{\text{ LCL, UCL }\} = \left\{\frac{\left(\chi^2_{.99865,n-1}\right)\overline{s^2}}{n-1}, \frac{\left(\chi^2_{.00135,n-1}\right)\overline{s^2}}{n-1}\right\}$$

Previous video:

$$P\left(\bar{X} - 3\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 3\frac{\sigma}{\sqrt{n}}\right) = .9973$$

$$P\left(\boxed{\mu - 3\frac{\sigma}{\sqrt{n}}} < \bar{X} < \boxed{\mu + 3\frac{\sigma}{\sqrt{n}}}\right) = .9973$$

μ is unknown and will be estimated with the average among the sample means during the Phase 1 base period. σ is unknown and will be estimated with the square root of the average among the sample variances.

$$P\left(\boxed{\mu - 3\frac{\sigma}{\sqrt{n}}} < \bar{X} < \boxed{\mu + 3\frac{\sigma}{\sqrt{n}}}\right) = .9973$$

μ is unknown and will be estimated with the average among the sample means during the Phase 1 base period. σ is unknown and will be estimated with the square root of the average among the sample variances.

$\bar{X}$ chart control limits:

$$\{\text{ LCL, UCL }\} = \left\{\bar{\bar{X}} - 3\sqrt{\frac{\overline{s^2}}{n}}, \bar{\bar{X}} + 3\sqrt{\frac{\overline{s^2}}{n}}\right\}$$

	LCL	UCL
$\bar{X}$ chart	$\bar{\bar{X}} - 3\sqrt{\frac{\overline{s^2}}{n}}$	$\bar{\bar{X}} + 3\sqrt{\frac{\overline{s^2}}{n}}$
s^2 chart	$\frac{(\chi^2_{.99865,n-1})\overline{s^2}}{n-1}$	$\frac{(\chi^2_{.00135,n-1})\overline{s^2}}{n-1}$

$\bar{\bar{X}}$ = average among the sample means during Phase 1

$\overline{s^2}$ = average among the sample variances during Phase 1

m = # of samples in Phase 1 n = # of observations per sample

Don’t confuse m and n !

	LCL	UCL
$\bar{X}$ chart	$\bar{\bar{X}} - 3\sqrt{\frac{\overline{s^2}}{n}}$	$\bar{\bar{X}} + 3\sqrt{\frac{\overline{s^2}}{n}}$
s^2 chart	$\frac{\left(\chi^2_{.99865,n-1}\right)\overline{s^2}}{n-1}$	$\frac{\left(\chi^2_{.00135,n-1}\right)\overline{s^2}}{n-1}$

<u>Assumptions</u>:

Data is normally distributed.

The process variance is always constant. Therefore any alarms on the s^2 chart indicate the $\bar{X}$ chart is useless and should be discarded.

Process is in control during Phase 1.

Sweet Club Inc. is a successful manufacturer of baseball bats. Most baseball bats are made of ash, but Sweet Club specializes in maple bats. Many players prefer maple because it is a harder wood than ash. However, maple bats are controversial because they tend to shatter into pieces, whereas ash bats tend to just crack. Sweet Club Inc. has hired you (an engineer with statistics expertise) to produce 5 control charts for various processes at their facility.

Charts C & D

$\overline{X}$ and s^2 chart

Since the hardness of the maple bats contributes to their popularity, Sweet Club Inc. would like to closely monitor the density of the bats they produce. Of the dozens of bats produced per day, five are selected at random and the sample mean density is measured (mg/cm^3). You have been provided with data for the last 88 days.

Use the first $m = 22$ samples as the Phase 1 base period.

Charts C & D

Maple bat density (mg/cm^3).

A sample of n = 5 is taken each day.

88 samples total.

The first $m = 22$ samples is the Phase 1 base period.

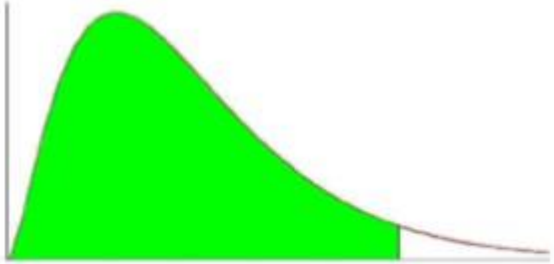
$$\bar{\bar{X}} \approx 700.000 \qquad \overline{s^2} \approx 24.044$$

center line = $\bar{\bar{X}} \approx 700.000$

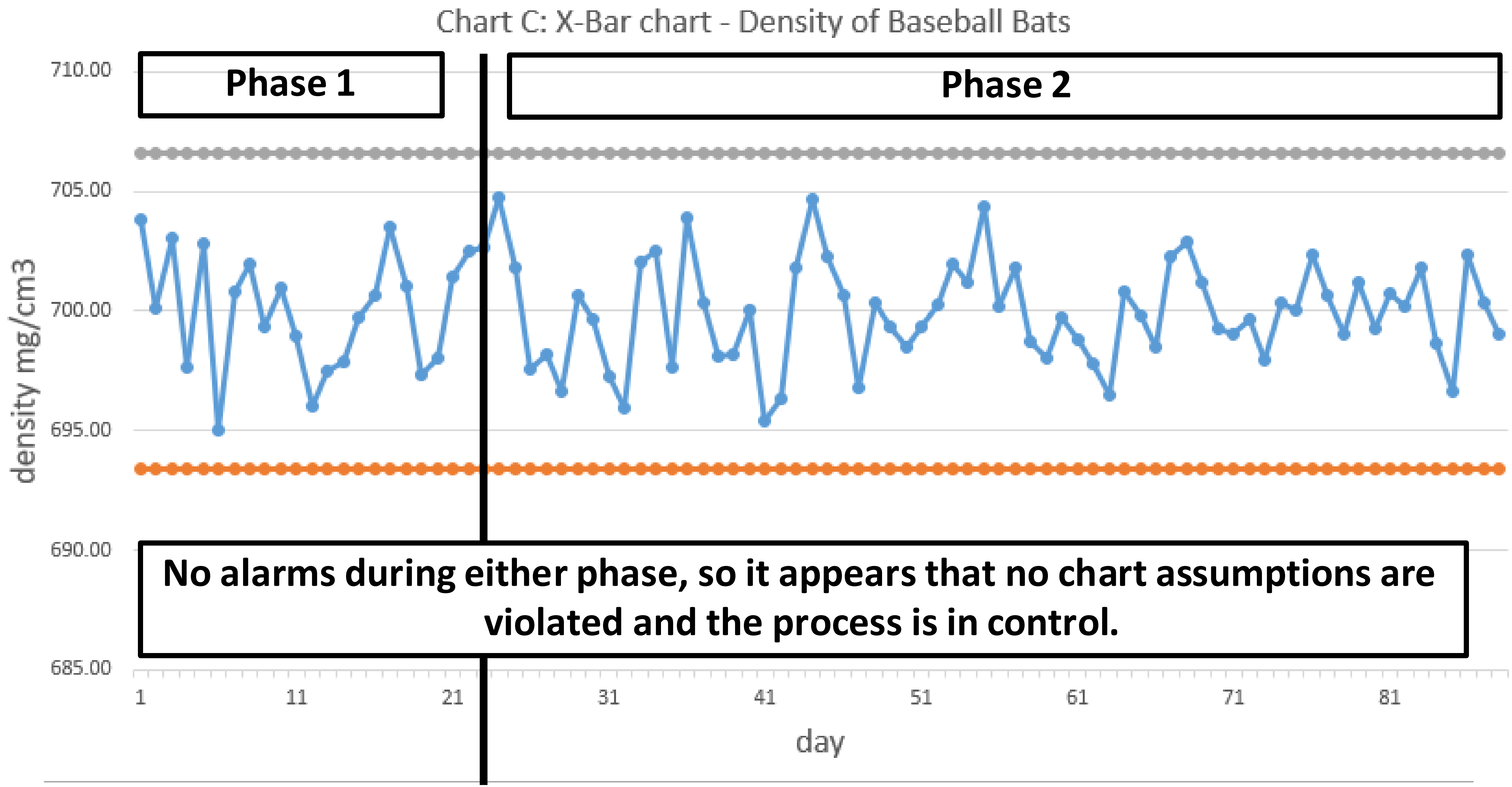
$\overline{\overline{X}} \approx 700.000 \qquad \overline{s^2} \approx 24.044 \qquad n = 5$

	LCL	UCL
$\overline{X}$ chart	$\overline{\overline{X}} - 3\sqrt{\frac{\overline{s^2}}{n}} = 700 - 3\sqrt{\frac{24.044}{5}} \approx 693.421$	$\overline{\overline{X}} + 3\sqrt{\frac{\overline{s^2}}{n}} = 700 + 3\sqrt{\frac{24.044}{5}} \approx 706.579$
s^2 chart	$\frac{\left(\chi^2_{.99865,\,n-1}\right)\overline{s^2}}{n-1} = \frac{(0.106)(24.044)}{4} \approx 0.637$	$\frac{\left(\chi^2_{.00135,\,n-1}\right)\overline{s^2}}{n-1} = \frac{(17.8)(24.044)}{4} \approx 106.995$

Chi-Square Table



df	0.99865	0.95	0.90	0.10	0.05	0.025	0.01	0.00135
1	0.000	0.004	0.016	2.706	3.841	5.024	6.635	10.273
2	0.003	0.103	0.211	4.605	5.991	7.378	9.210	13.215
3	0.030	0.352	0.584	6.251	7.815	9.348	11.345	15.630
4	$\boxed{0.106}$	0.711	1.064	7.779	9.488	11.143	13.277	$\boxed{17.800}$
5	0.238	1.145	1.610	9.236	11.070	12.833	15.086	19.821
6	0.423	1.635	2.204	10.645	12.592	14.449	16.812	21.739



Michigan Tech

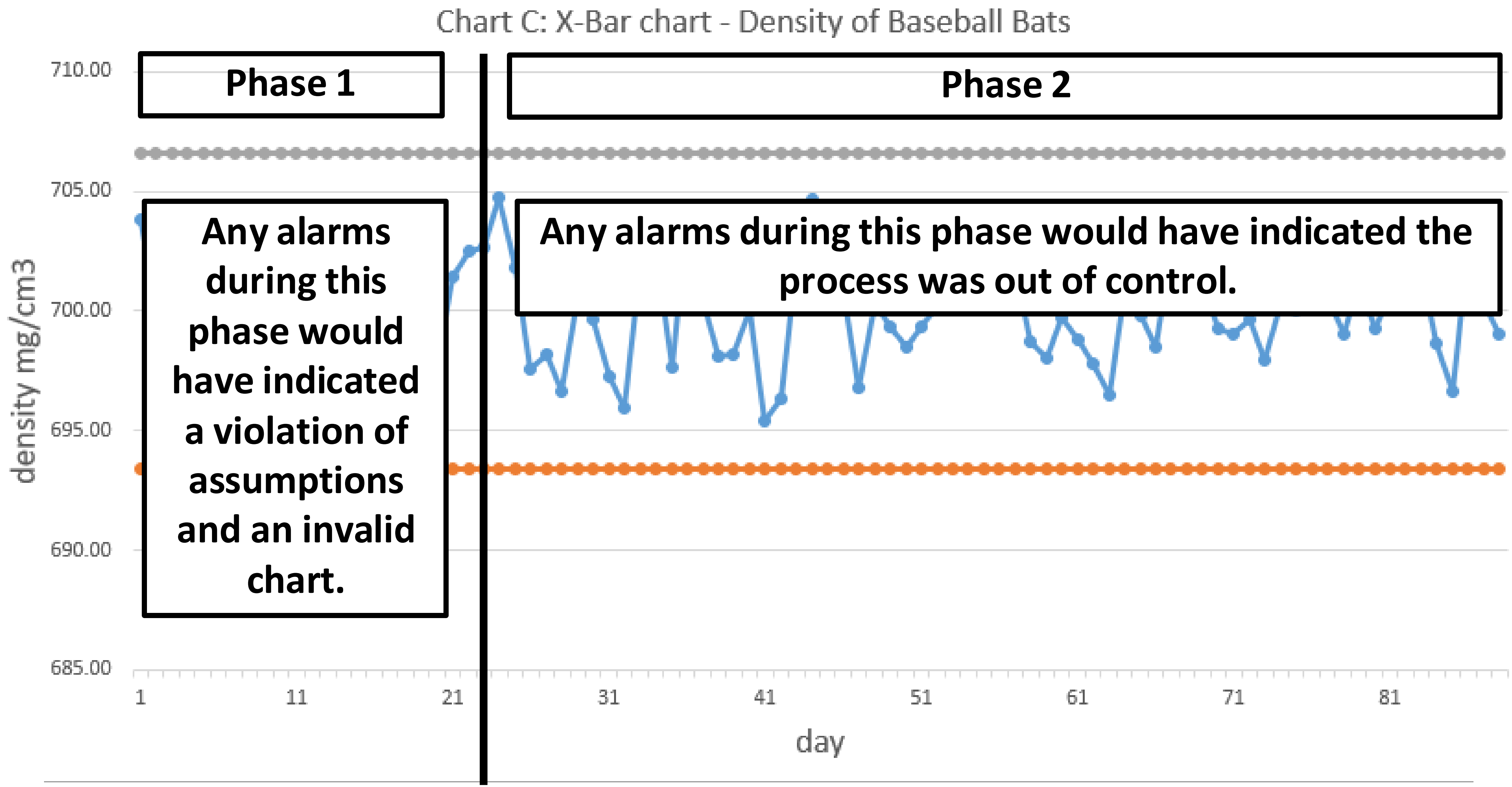


Chart C: X-Bar chart - Density of Baseball Bats
Phase 1
Phase 2
Any alarms during this phase would have indicated a violation of assumptions and an invalid chart.
Any alarms during this phase would have indicated the process was out of control.
710.00
705.00
700.00
695.00
690.00
685.00
density mg/cm3
1
11
21
31
41
51
61
71
81
day

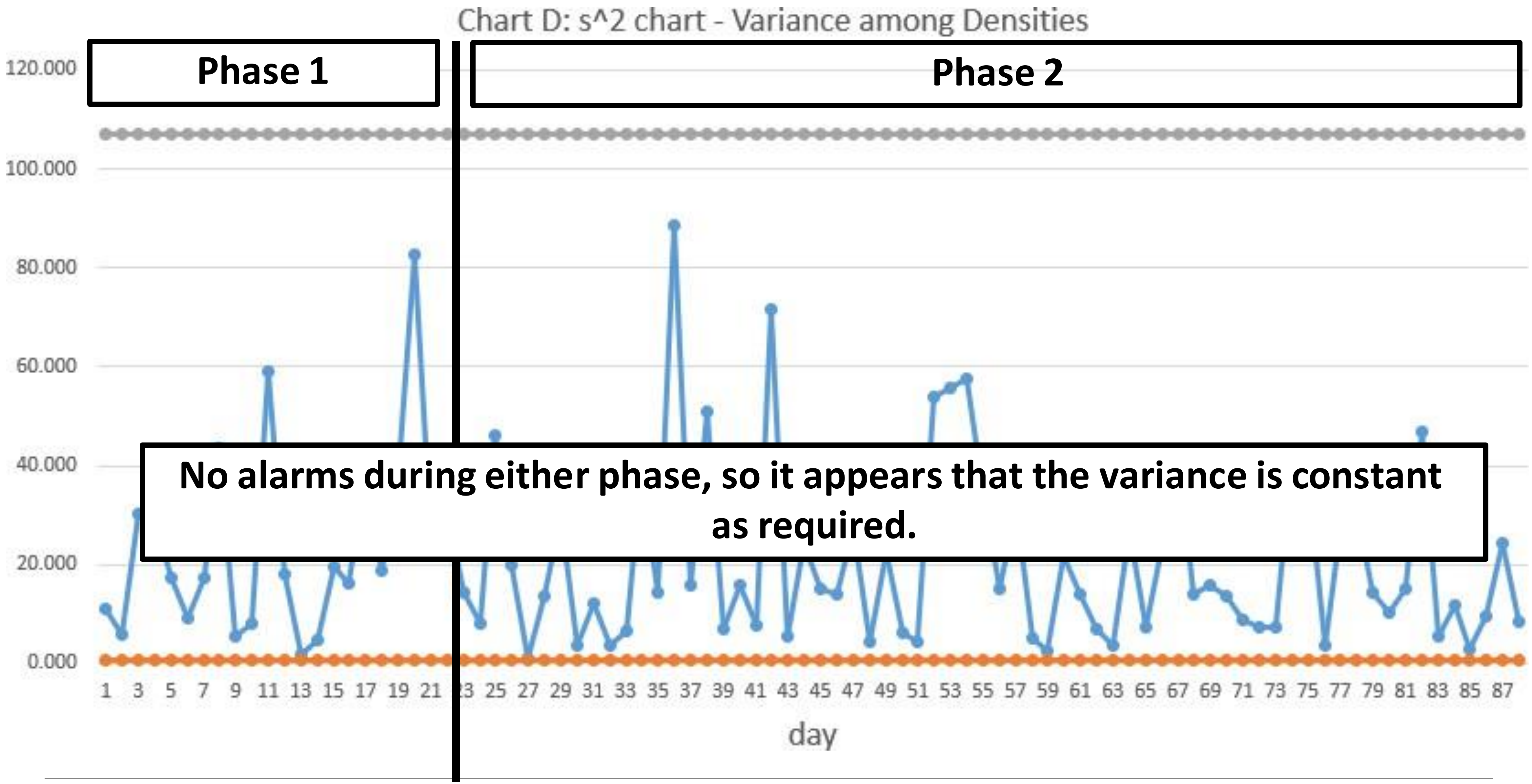
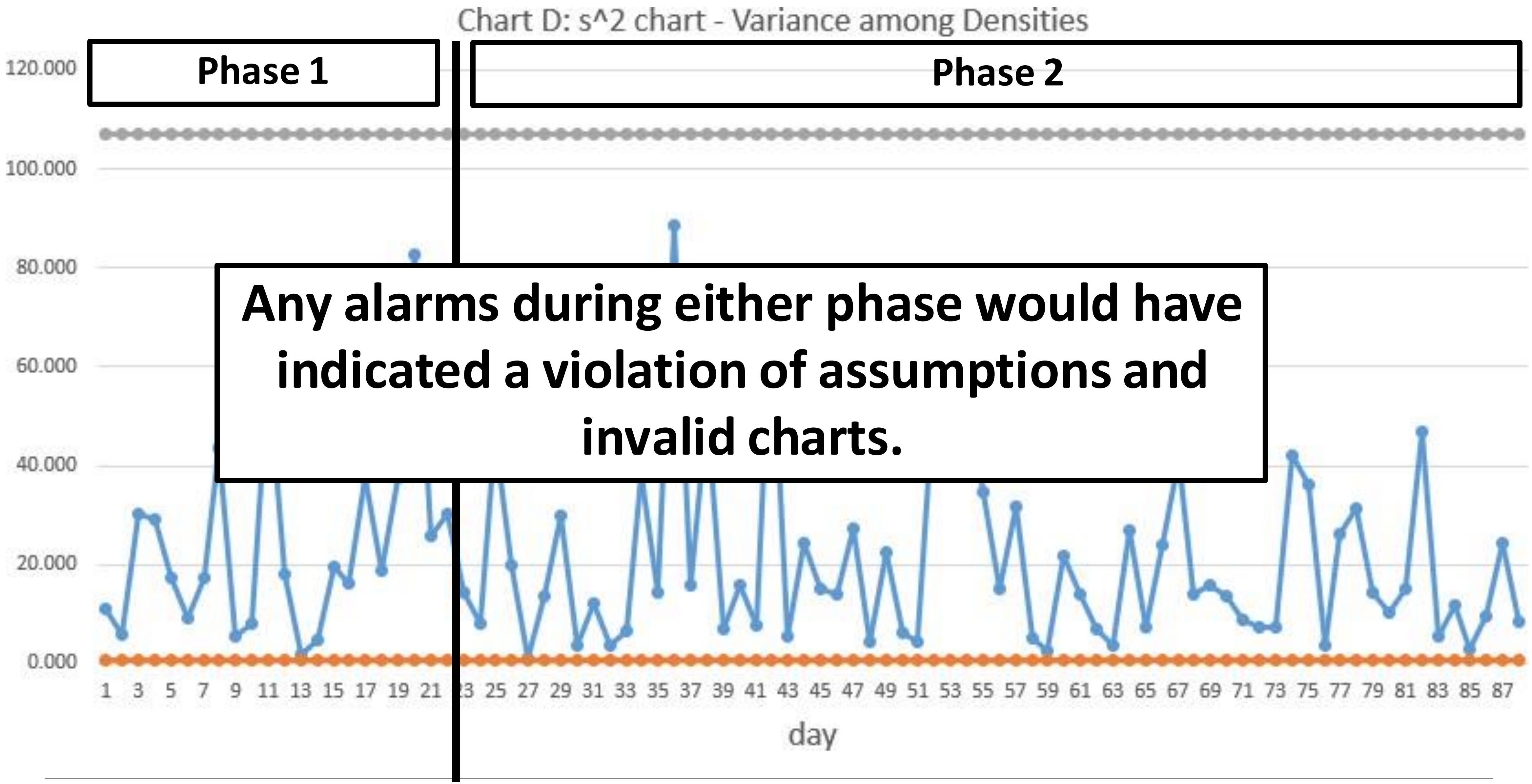


Chart D: s^2 chart - Variance among Densities
Phase 1
Phase 2
120.000
100.000
80.000
60.000
40.000
20.000
0.000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87
day
No alarms during either phase, so it appears that the variance is constant as required.



Any alarms during either phase would have indicated a violation of assumptions and invalid charts.

Engineering Statistics

Chapter 5: control charts

Video 12.03

X chart

Department of Mathematical Sciences

Philip Kendall, Senior Lecturer

Michigan Tech

An $\overline{X}$ chart monitored repeated observations of $\overline{X}$ from normally distributed random samples of size n

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

However due to the nature of the process it may be impossible, impractical, or simply too expensive to have samples of size greater than one.

When $n = 1$, there are no sample variances!

$$\cancel{s^2 = \frac{\sigma\left(x - \overline{X}\right)^2}{n-1} = \frac{\sigma\, x^2 - \frac{(\sigma\, x)^2}{n}}{n-1}}$$

An X chart will monitor single observations of X

$$X \sim N(\mu, \sigma)$$

The process variability will be measured with the moving range = MR

Observations: $X_1, X_2, X_3, \ldots$

Base period: first m observations

If there are m observations i[redacted] the base period, there will b[redacted] $m-1$ moving ranges.

$$MR_i = |X_{i+1} - X_i|$$

$$\overline{MR} = \frac{1}{m-1}\sum_{i=1}^{m-1} MR_i$$

[redacted]

i	[redacted]	[redacted]
1	[redacted]	5
2	2	0
[redacted]	[redacted]	[redacted]
5	3	7
6	10	

$$\overline{MR} = \frac{1}{5}(5+0+3+2+7) = \frac{17}{5} = \boxed{3.4}$$

Control Limits: $\mu_0 \pm 3\frac{\overline{MR}}{d_2}$

estimate with $\overline{X}$

$d_2 = 1.128$ when $n = 2$ (from a table)

$$\{\text{LCL, UCL}\} = \overline{X} \pm 2.66 \cdot \overline{MR}$$

<u>Assumptions</u>:

Data is normally distributed

Process is in control during base period

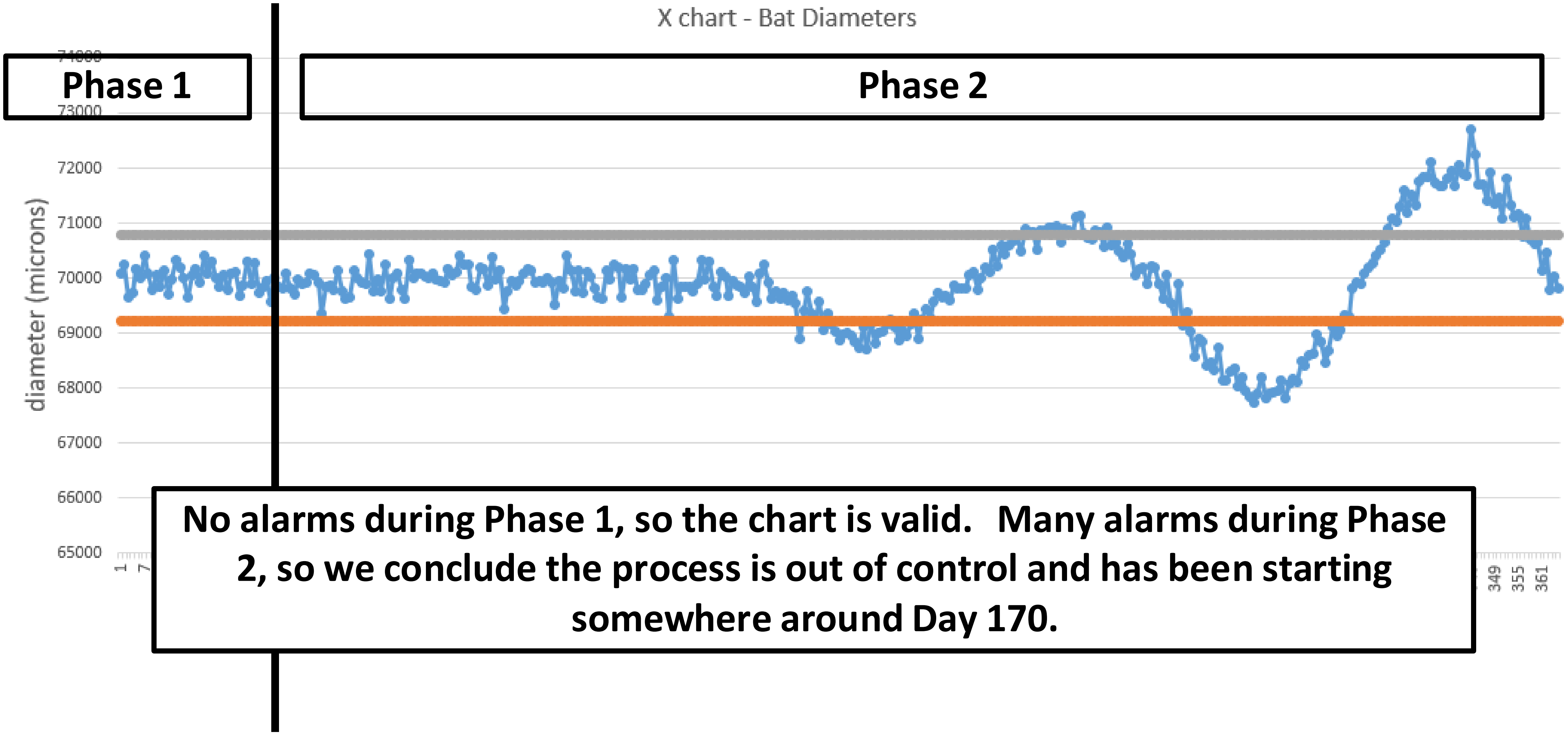
Chart E
X chart

An amateur baseball team purchases its maple bats from Sweet Club Inc. and requests that the diameter of the baseball bats at their thickest part be as close to 70 mm as possible (70mm = 70,000 microns). They are willing to pay an expensive price for their bats in exchange for a high level of accuracy. Once a day, Sweet Club Inc. randomly selects one bat and measures its diameter to the nearest micron (one millionth of a meter). Measuring the diameter to this accuracy is very expensive and is done with only one randomly selected bat per day. Data from the last 365 days has been provided. Use the first 40 observations as the base period.

$$m = 40 \qquad \overline{X} \approx 70000.00 \qquad \overline{MR} \approx 295.44$$

	LCL	UCL
X chart	$\overline{X} - 2.66 \cdot \overline{MR}$ $\approx 70000 - 2.66(295.44)$ $= 69214.12$	$\overline{X} + 2.66 \cdot \overline{MR}$ $\approx 70000 + 2.66(295.44)$ $= 70785.86$

center line = $\overline{X} \approx 70000.00$



No alarms during Phase 1, so the chart is valid. Many alarms during Phase 2, so we conclude the process is out of control and has been starting somewhere around Day 170.

Engineering Statistics

Chapter 5: **control charts**

Video 12.04
np chart

Department of Mathematical Sciences
Philip Kendall, Senior Lecturer

Michigan Tech

An **np chart** monitors the number of defective units in samples of size n

X_i = # of defective units in the ith sample

X_i is Binomial with p = P(defective) and n = # of trials

Recall: if $X \sim BIN(n, p)$ then

$$\mu = E(X) = np$$

$$\sigma^2 = Var(X) = np(1-p)$$

$$H_0: p = p_0 \rightarrow np = np_0$$
$$H_a: p \neq p_0 \rightarrow np \neq np_0$$

It is recommended to have $np_0 \geq 10$ and $n(1 - p_0) \geq 10$ so the LCL will be positive.

{ LCL, UCL } = center line $\pm$ 3·SD(observed statistic)

$$np_0 \pm 3\sqrt{np_0(1 - p_0)}$$

p_0 will be estimated with $\overline{p} =$ the average proportion of defective units during the base period.

$$np_0 \pm 3\sqrt{np_0(1-p_0)}$$

p_0 will be estimated with $\overline{p}$ = the average proportion of defective units during the base period.

$$\{\text{ LCL, UCL }\} = n\overline{p} \pm 3\sqrt{n\overline{p}(1-\overline{p})}$$

<u>Assumptions</u>:

Data is Binomial

Process is in control during base period

m = length of base period

n = units per sample

X_i = # of defective units in sample i

$$\overline{p} = \frac{1}{mn}\sum_{i=1}^{m} X_i$$

Chart F
np chart

Before baseball bats are shipped to retailers, some are randomly selected and subjected to the "Smash Test". To pass the Smash Test, a bat must endure several collisions with baseballs travelling 100 mph without shattering into pieces. Once a week, Sweet Club Inc. takes a random sample of 500 bats, subjects them to the Smash Test, and records the number of failures. Data from the last 52 weeks has been provided. Use the first 24 observations as the base period.

$m = 24$ [redacted] $= 500$ [redacted]

252 defectives du[redacted]se period

$\bar{p} = \frac{1}{mn}\sum_{i=1}^{m} X_i =$ [redacted] $24(500)$ [redacted]

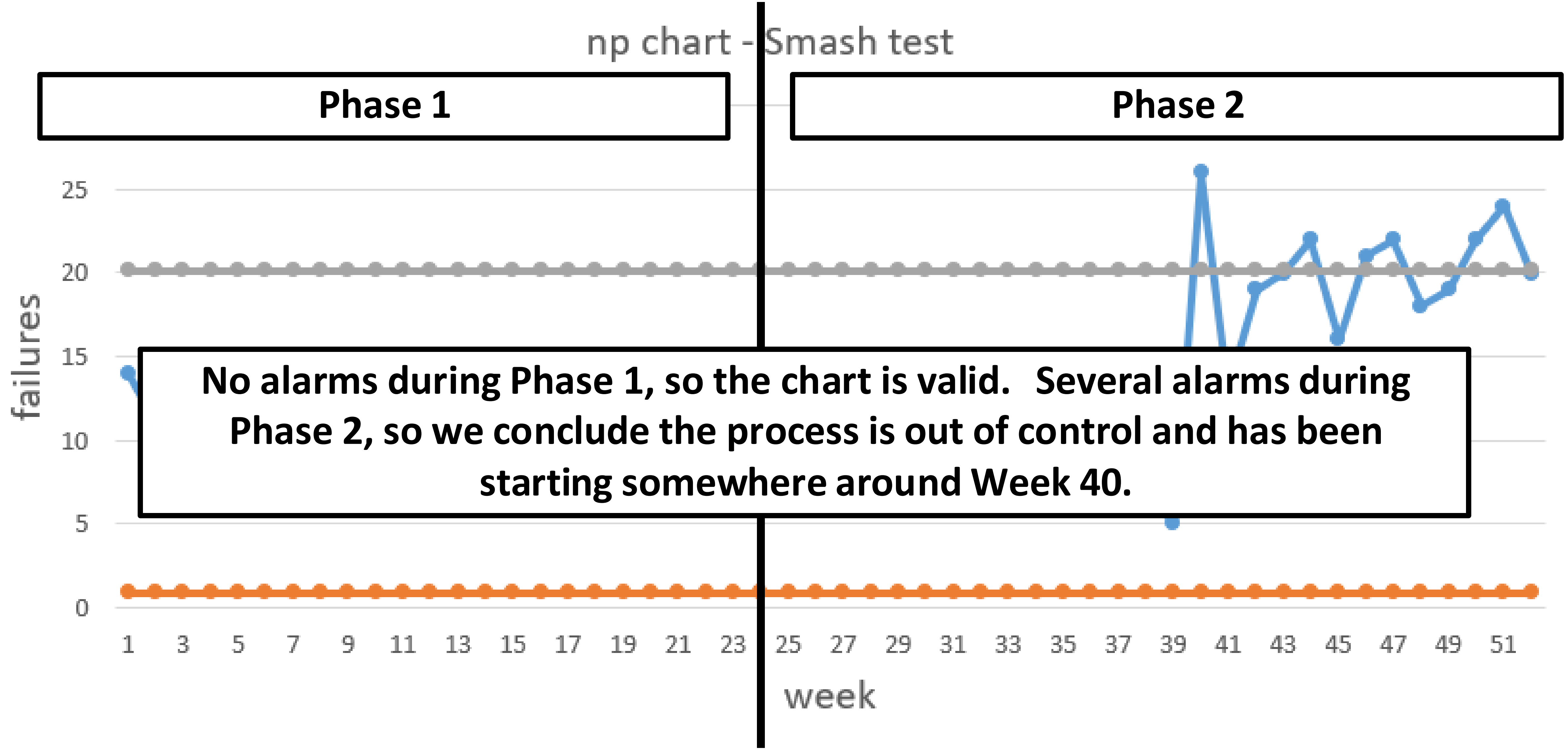
24 week base period

500 bats per week

center line = $n\bar{p} = 500(0.021) = 10.5$

$m = 24 \qquad n = 500 \qquad \bar{p} = 0.021$

	LCL	UCL
np chart	$n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} =$ $10.5 - 3\sqrt{500(0.021)(0.979)} =$ 0.882	$n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} =$ $10.5 + 3\sqrt{500(0.021)(0.979)} =$ 20.118



No alarms during Phase 1, so the chart is valid. Several alarms during Phase 2, so we conclude the process is out of control and has been starting somewhere around Week 40.

Introductory Statistics
with R

Unit 12: control charts

Video 12.05
c chart

Department of Mathematical Sciences
Philip Kendall, Senior Lecturer

Michigan Tech 1885

A **c chart** monitors the number of occurrences of an event in a specified time period

X_i = # of occurrences in the ith period

X_i is Poisson with λ = the mean rate of occurrences

Recall: If $X \sim POI(\lambda)$ then

$$\mu = E(X) = \lambda$$

$$\sigma^2 = Var(X) = \lambda$$

$$H_0 : \lambda = \lambda_0$$
$$H_a : \lambda \neq \lambda_0$$

It is recommended to have $\lambda_0 \geq 10$ so the LCL will be positive.

{ LCL, UCL } = center line $\pm$ 3·SD(observed statistic)

$$\lambda_0 \pm 3\sqrt{\lambda_0}$$

λ_0 will be estimated with $\bar{c}$ = the average count of occurrences during the base period.

$$\lambda_0 \pm 3\sqrt{\lambda_0}$$

λ_0 will be estimated with $\bar{c}$ = the average count of occurrences during the base period.

$$\bar{c} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

$$\{\ \text{LCL, UCL}\ \} = \bar{c} \pm 3\sqrt{\bar{c}}$$

<u>Assumptions</u>:

Data is Poisson

Process is in control during base period

Chart G
c chart

When pieces of maple first arrive at Sweet Club Inc., they are placed on a lathe and shaped into the form of a baseball bat. Sweet Club Inc. has a few dozen lathes in operation at all times. An engineer working at the facility records how many times per month a lathe breaks down and needs to be repaired. Sweet Club Inc. would like to monitor how often repairs are needed. Data from the last 105 months has been provided. Use the first 30 observations as the base period.

$$m = 30$$

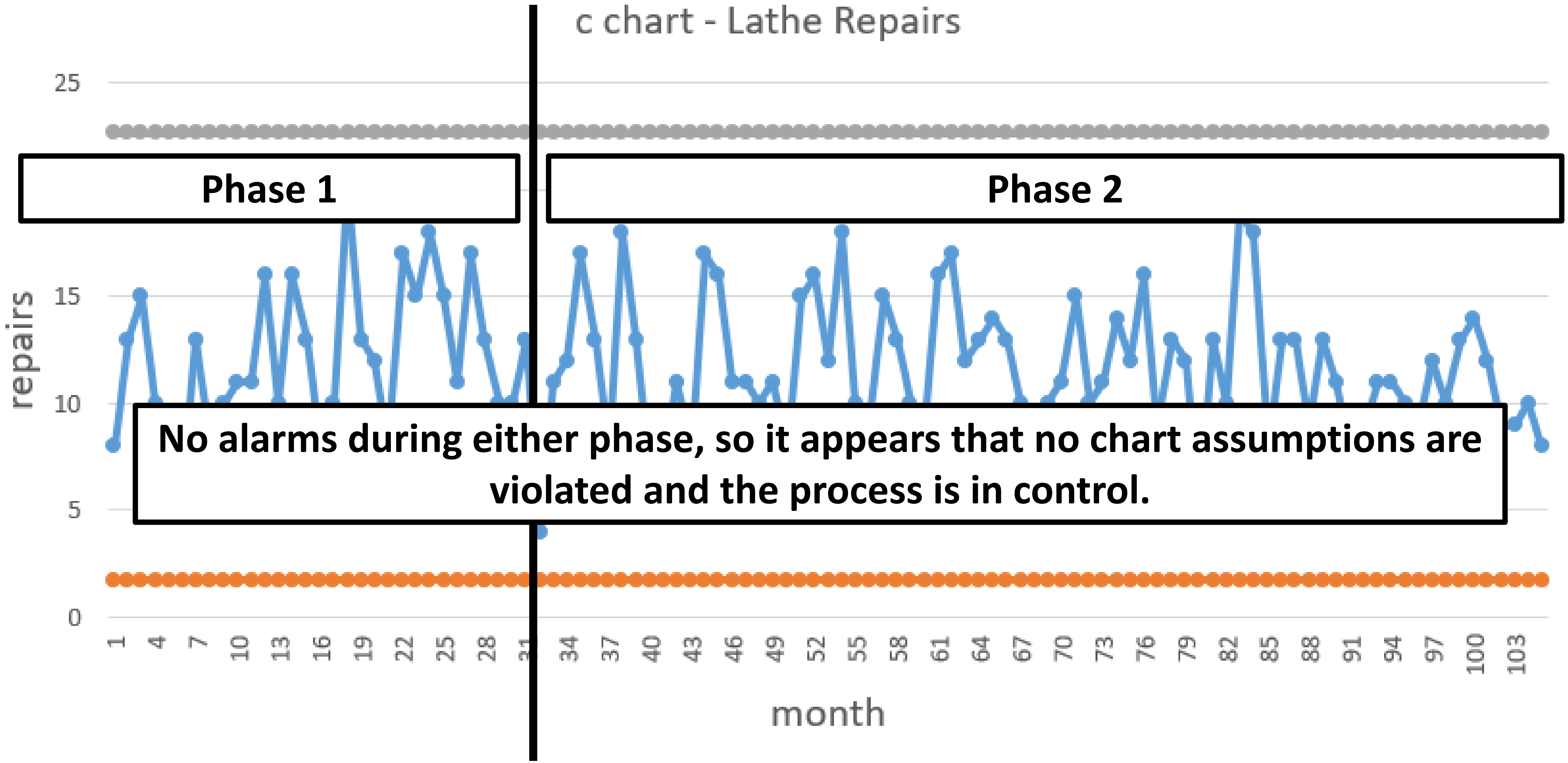
366 repairs during base period

$$\bar{c} = \frac{1}{m}\sum_{i=1}^{m} X_i = \frac{366}{30} = 12.2$$

center line = $\bar{c} = 12.2$

$$m = 30 \qquad \bar{c} = 12.2$$

	LCL	UCL
c chart	$\bar{c} - 3\sqrt{\bar{c}}$ $12.2 - 3\sqrt{12.2}$ 1.721	$\bar{c} + 3\sqrt{\bar{c}}$ $12.2 + 3\sqrt{12.2}$ 22.679



No alarms during either phase, so it appears that no chart assumptions are violated and the process is in control.

Introductory Statistics
with R

Unit 12: control charts

Video 12.06
final thoughts on control charts

Department of Mathematical Sciences
Philip Kendall, Senior Lecturer

Michigan Tech 1885

Final thoughts on Control Charts

Control Limit formulas

reactions to alarms

Control Limits vs. Specification Limits

Average Run Length = ARL

Other options for alarm rules exist.

Different types of charts exist.

Chart	LCL	UCL
$\overline{X}$	$\overline{\overline{X}} - 3\sqrt{\frac{\overline{s^2}}{n}}$	$\overline{\overline{X}} + 3\sqrt{\frac{\overline{s^2}}{n}}$
s^2	$\frac{\left(\chi^2_{n-1,0.99865}\right)\overline{s^2}}{n-1}$	$\frac{\left(\chi^2_{n-1,0.001355}\right)\overline{s^2}}{n-1}$
X	$\overline{X} - 2.66 \cdot \overline{MR}$	$\overline{X} + 2.66 \cdot \overline{MR}$
np	$n\overline{p} - 3\sqrt{n\overline{p}(1-\overline{p})}$	$n\overline{p} + 3\sqrt{n\overline{p}(1-\overline{p})}$
c	$\overline{c} - 3\sqrt{\overline{c}}$	$\overline{c} + 3\sqrt{\overline{c}}$

<table>
<tr><th>chart</th><th>assumptions</th><th>Phase 2 alarm?</th><th>Phase 1 alarm?</th></tr>
<tr><td>$\overline{X}$</td><td rowspan="2">*Process variance is constant
*in control during Phase 1
*sample means are normally distributed</td><td>Evidence that process is out of control.</td><td rowspan="5">Evidence that assumption(s) are violated – chart is invalid. Start over.</td></tr>
<tr><td>s^2</td><td>Evidence that assumption(s) are violated – chart is invalid. Start over.</td></tr>
<tr><td>X</td><td>*in control during Phase 1
*data is normally distributed</td><td rowspan="3">Evidence that process is out of control.</td></tr>
<tr><td>np</td><td>*in control during Phase 1
*data is Binomial</td></tr>
<tr><td>c</td><td>*in control during Phase 1
*data is Poisson</td></tr>
</table>

Control Limits vs. Specification Limits

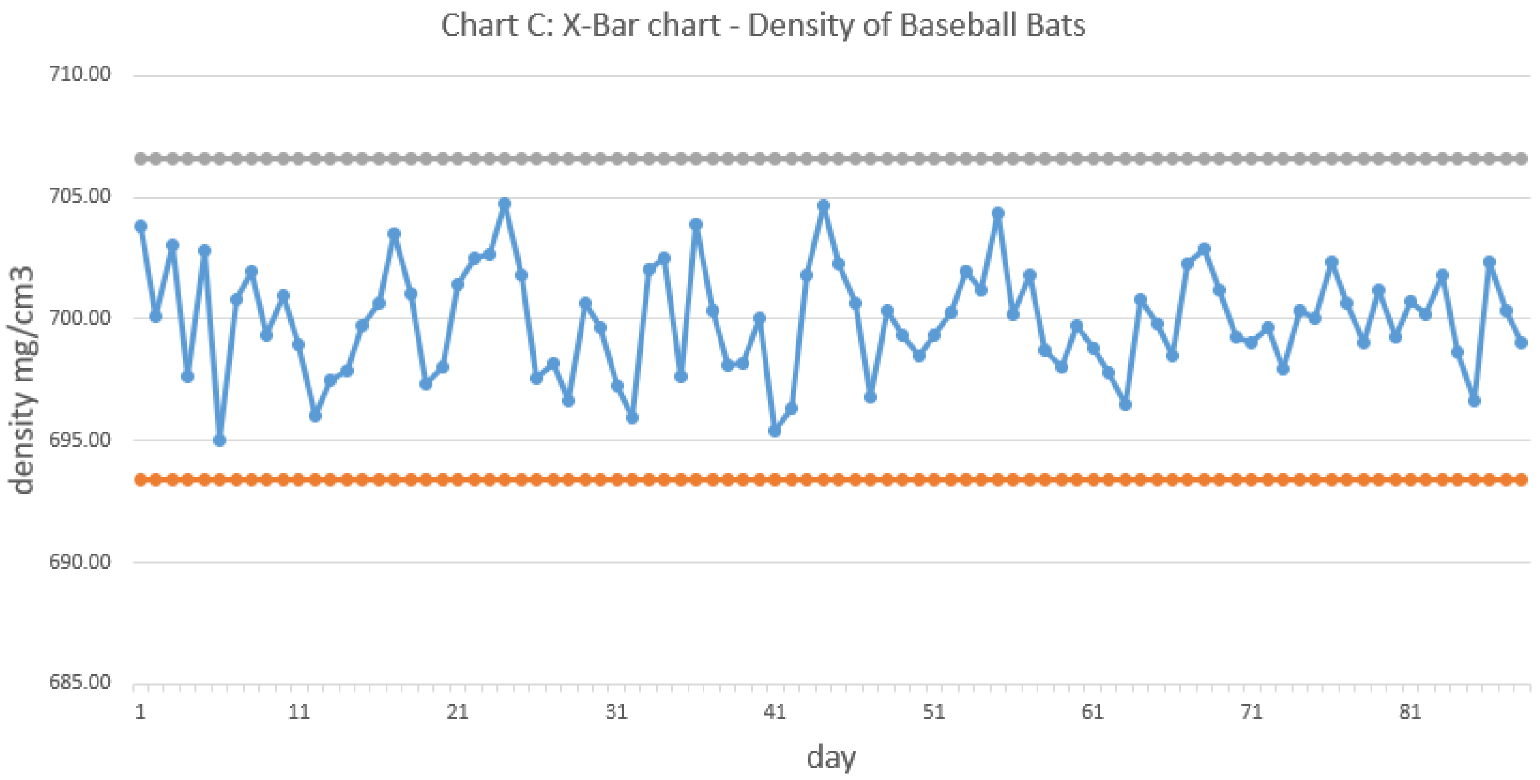
Control Limits { LCL, UCL } are sometimes called the "voice of the process", are computed with sample data, and are used to assess whether or not a process is stable. Based on the data collected during the Phase 1 base period, the LCL and UCL reflect a range of variation we should naturally expect from an in-control process.

On the other hand, Specification Limits { LSL, USL } are sometimes called the "voice of the customer" and define an acceptable range for the product. They may be artificially imposed on the process through regulations or other factors.

For example, consider the earlier example (Chart C) where we drew an $\overline{X}$ chart to monitor the density of maple baseball bats.

LCL $\approx$ 693 and UCL $\approx$ 707

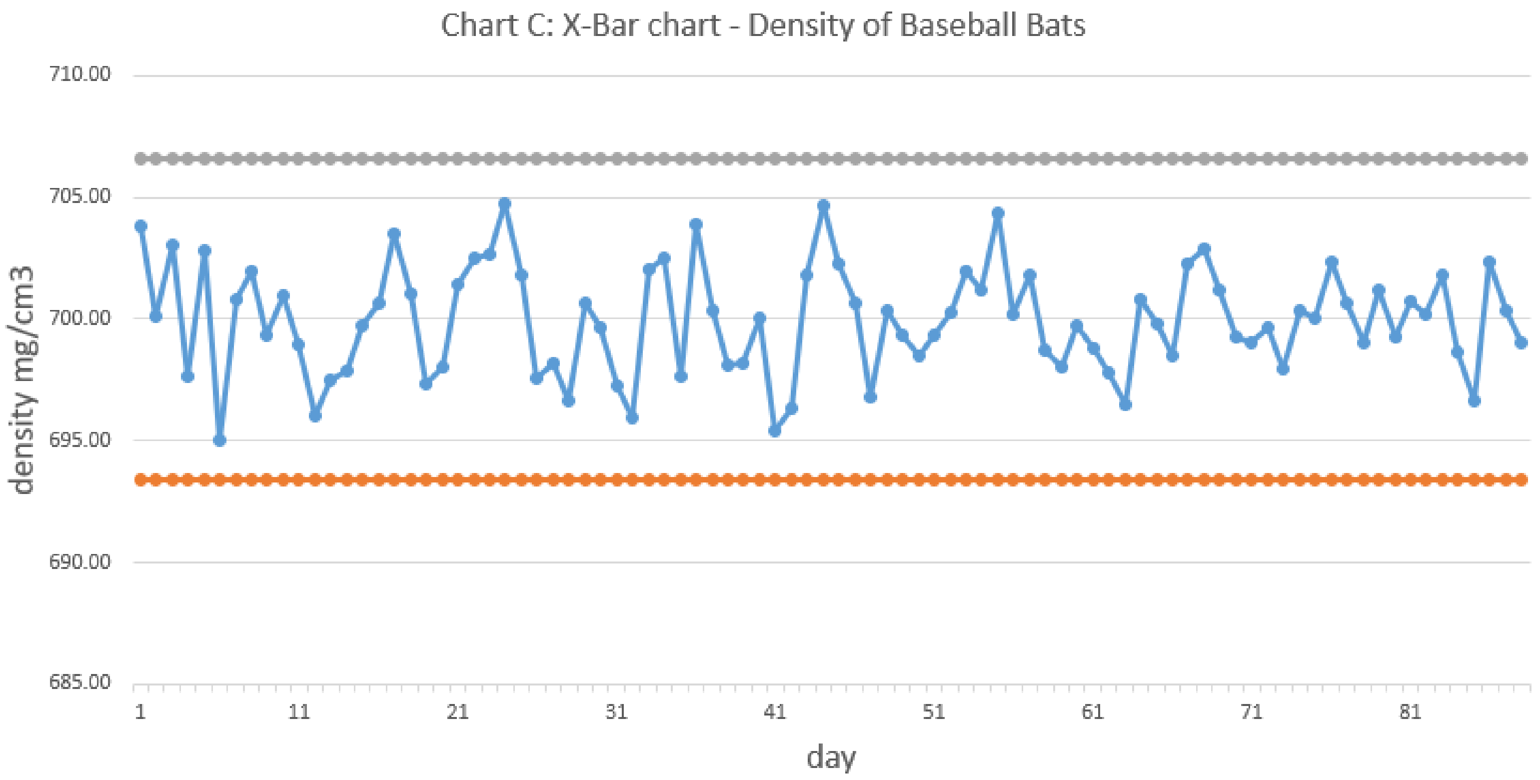
The $\overline{X}$ chart was fine with no alarms during either phase.



Michigan Tech

LCL ≈ 693 and UCL ≈ 707

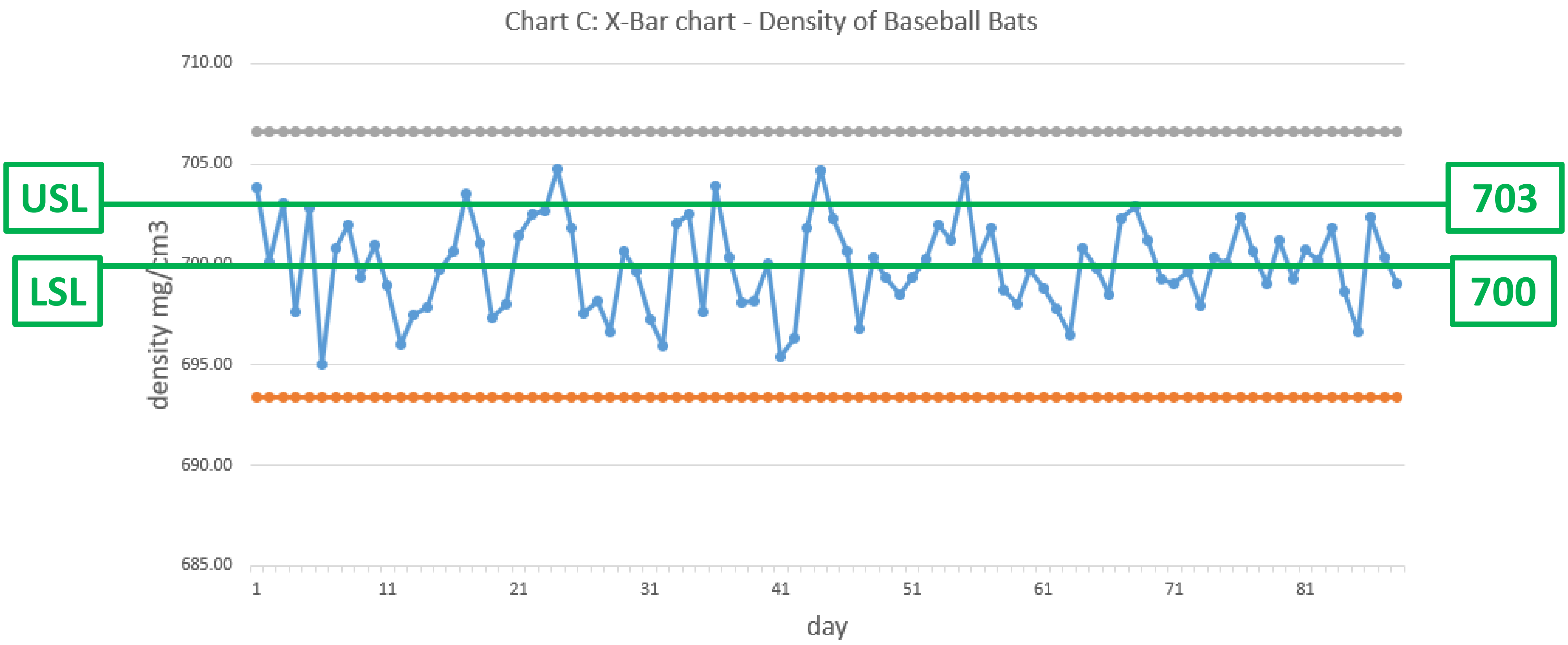
But what if due to customer demands or some other form of regulation, it was determined that only densities between 700 and 703 were acceptable?



Michigan Tech

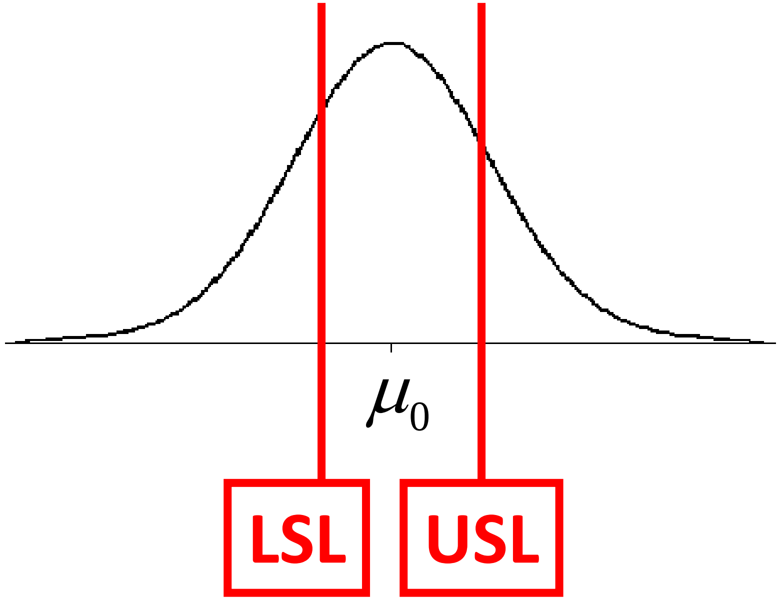
LCL ≈ 693 and UCL ≈ 707, but LSL = 700 and USL = 703

Many observed sample means were outside of that range!

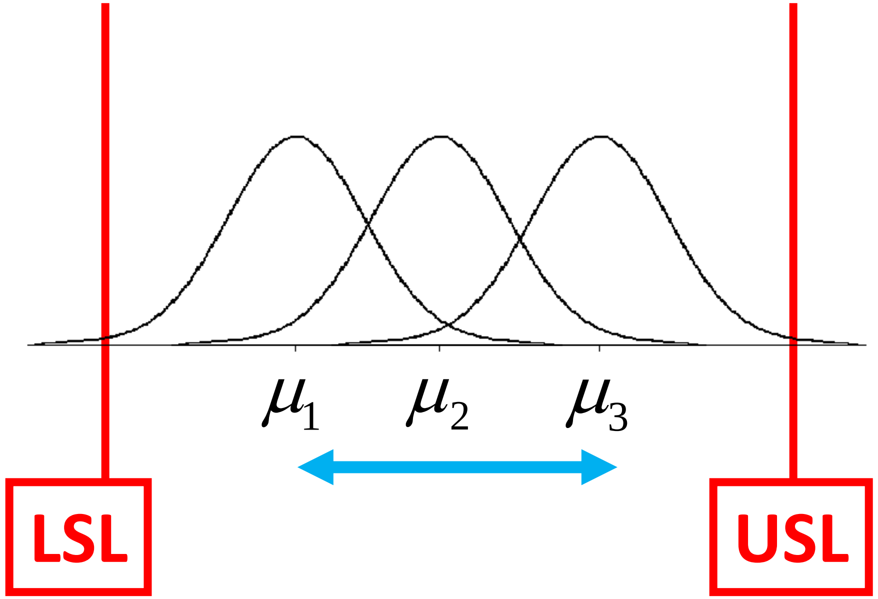


Note: it is possible for an in-control process to regularly produce observations outside of specification limits and vice-versa.

Process is in control but many observations are outside of specification limits.



Process is out of control but almost all observations are within specification limits.



Average Run Length = ARL

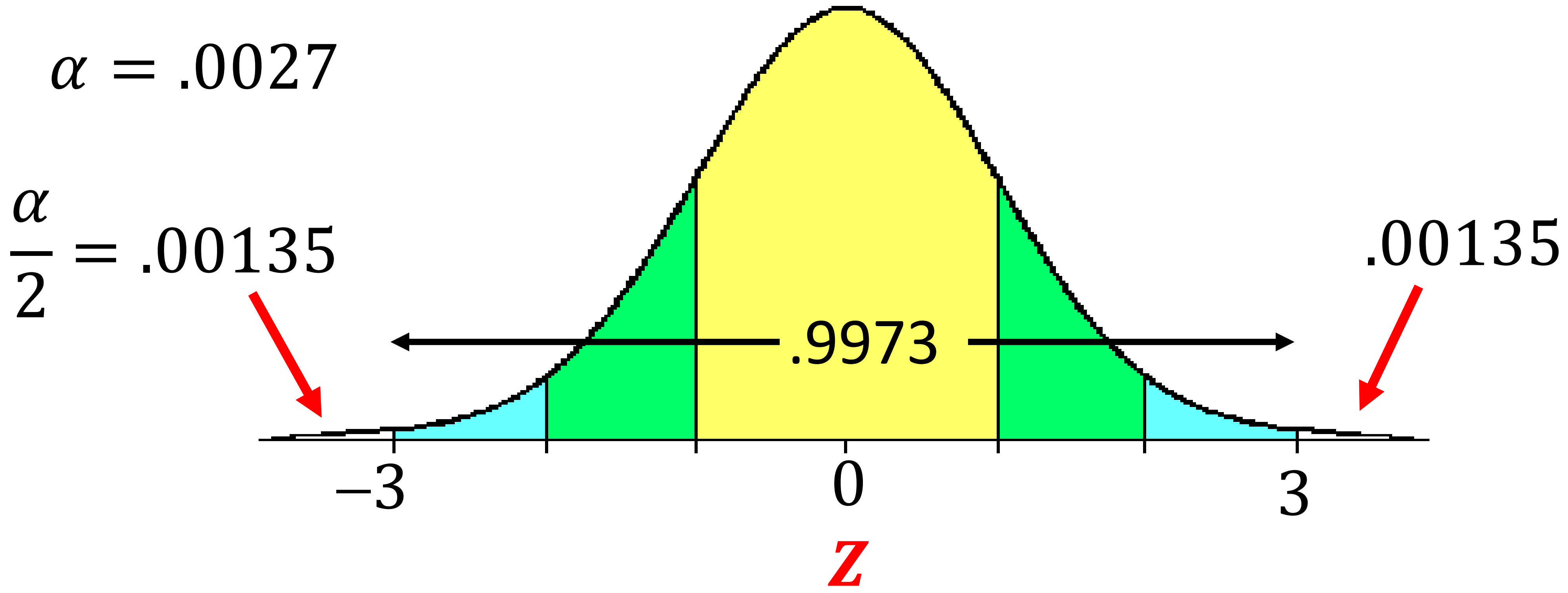
Define a random variable N = the # of samples needed to observe the first alarm.

N is a **Geometric** random variable, which counts the number of trials needed to observe the first "success".

$$\text{P(success)} = p$$

If $X \sim GEO(p)$ then $E(X) = \frac{1}{p}$

$$E(N) = \frac{1}{P(\text{alarm})}$$



$$E(N) = \frac{1}{P(\text{alarm})} = \frac{1}{.0027} \approx 370$$

Recall: an in-control process using $\alpha = .0027$ will false alarm about once in every 370 samples.

So under these conditions, **ARL = 370**.

Theoretical example.

Suppose σ is known to be 16 and $\mu_0 = 100$

If n = 4, the LCL and UCL are:

$$\mu_0 \pm 3\frac{\sigma}{\sqrt{n}} = 100 \pm 3\left(\frac{16}{\sqrt{4}}\right) = 100 \pm 24 = \{76{,}124\}$$

If n = 9, the LCL and UCL are:

$$\mu_0 \pm 3\frac{\sigma}{\sqrt{n}} = 100 \pm 3\left(\frac{16}{\sqrt{9}}\right) = 100 \pm 16 = \{84{,}116\}$$

Theoretical example.
Suppose σ is known to be 16 and $\mu_0 = 100$

If μ actually equals 116 (the actual mean is one standard deviation greater than the ideal mean), on average how long will we wait for the chart to alarm and tell us the process is out of control?

Clearly, if the process is out of control we would like $E(N) = \frac{1}{P(\text{alarm})}$ to be small.

$$\boxed{\text{If } n = 4 \text{ then LCL} = 76 \text{ and UCL} = 124}$$

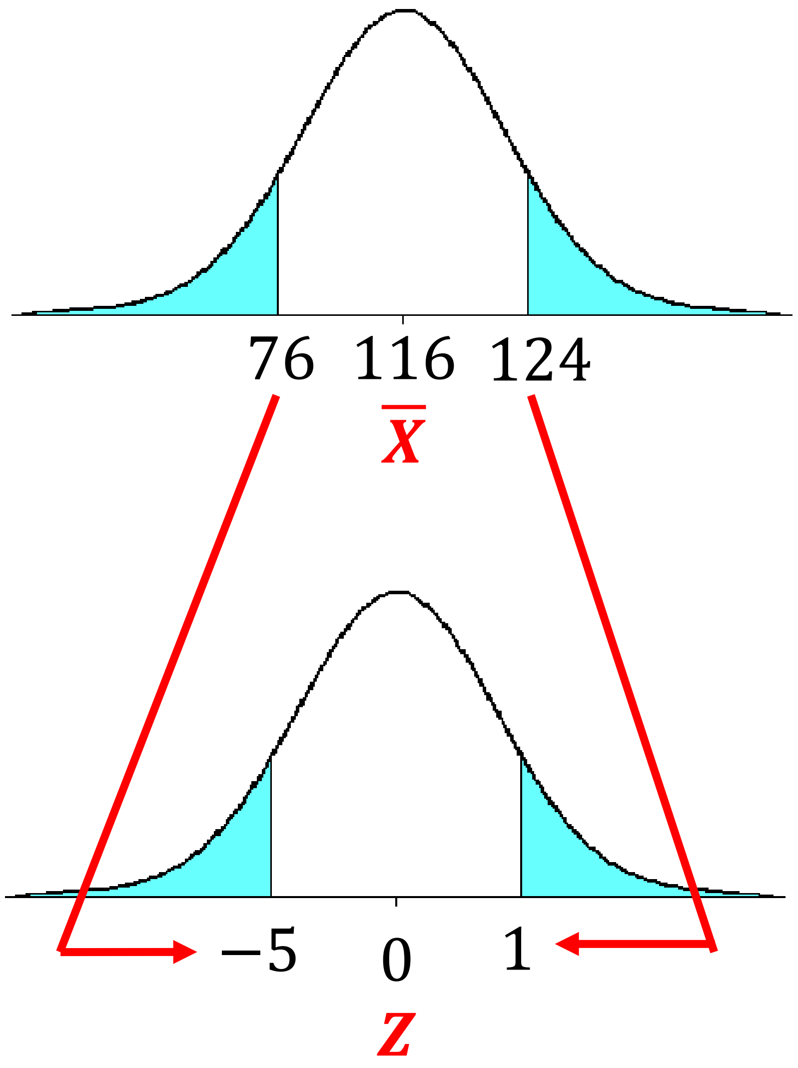
P(signal) = 1 – P(no signal)

$$= 1 - P(76 < \bar{X} < 124)$$

$$= 1 - P\left(\frac{76 - 116}{16/\sqrt{4}} < Z < \frac{124 - 116}{16/\sqrt{4}}\right)$$

$$= 1 - P(-5 < Z < 1)$$

$$= 1 - [.8413 - (\approx 0)] = .1587$$



$\boxed{\text{If } n = 9 \text{ then LCL} = 84 \text{ and UCL} = 116}$

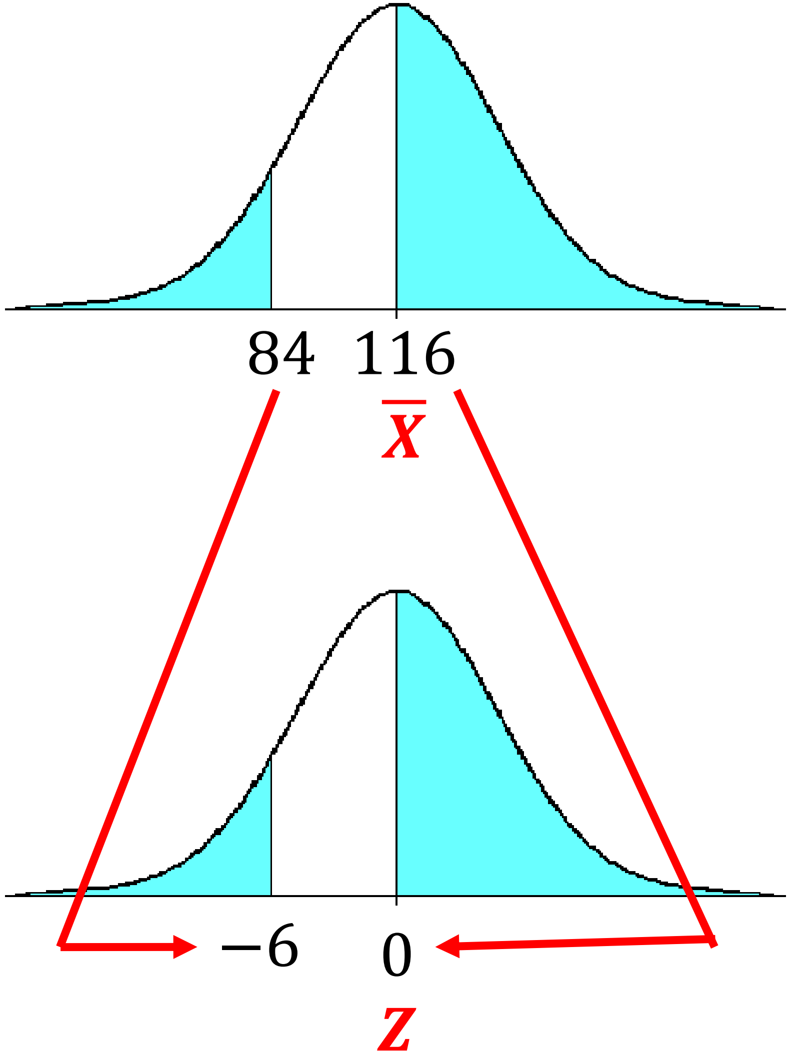
P(signal) = 1 – P(no signal)

$$= 1 - P(84 < \overline{X} < 116)$$

$$= 1 - P\left(\frac{84 - 116}{16/\sqrt{9}} < Z < \frac{116 - 116}{16/\sqrt{9}}\right)$$

$$= 1 - P(-6 < Z < 0)$$

$$= 1 - [.50 - (\approx 0)] = .50$$



Michigan Tech

Theoretical example.

Suppose σ is known to be 16 and $\mu_0 = 100$

if $n = 4$

actual μ	$P(\text{alarm})$	ARL
100	.0027	≈ 370
116	.1587	≈ 6.30
132	.8413	≈ 1.19

if $n = 9$

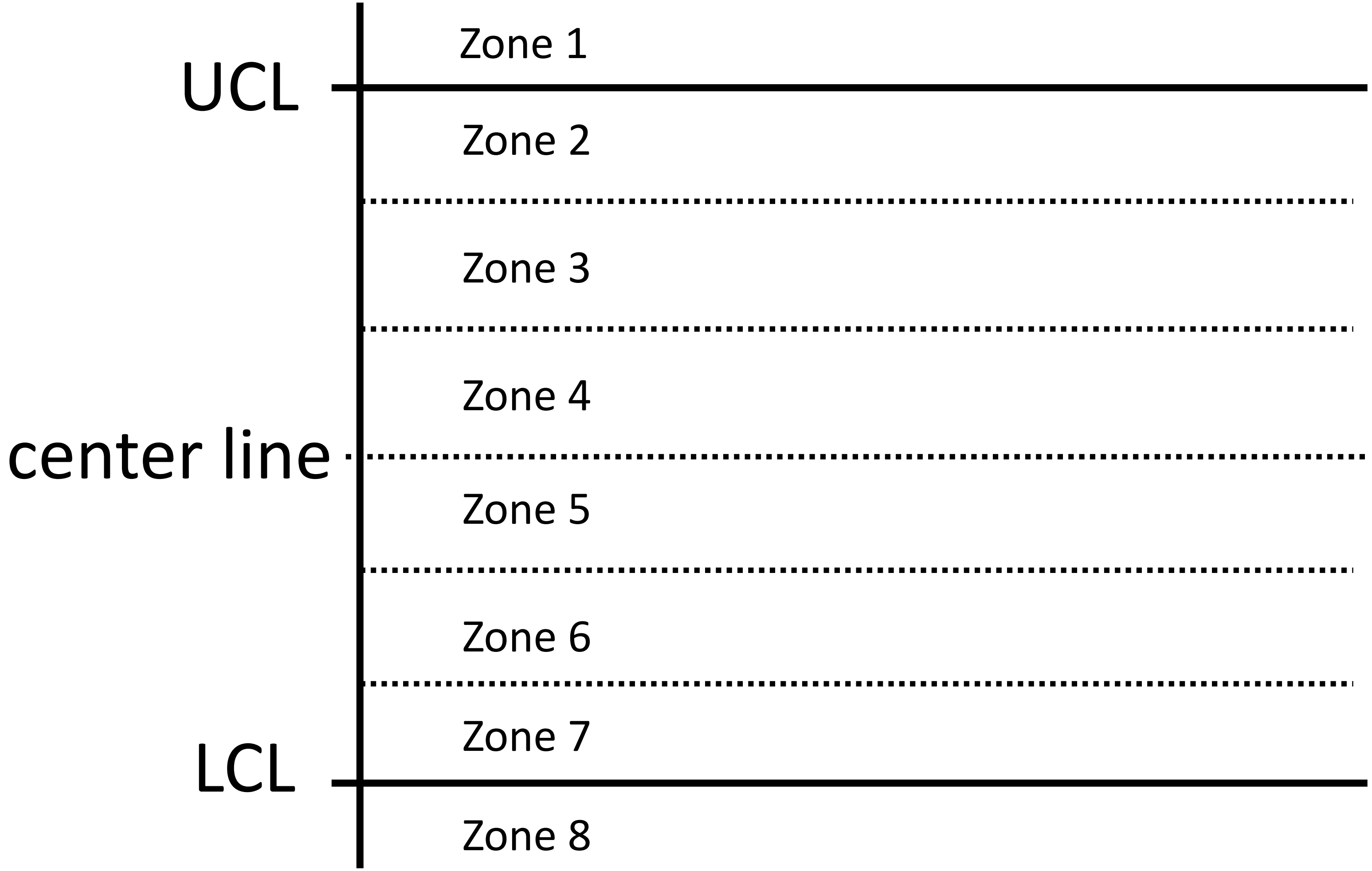
actual μ	$P(\text{alarm})$	ARL
100	.0027	≈ 370
116	.50	2
132	≈ 1	≈ 1

ARL = average run length is lower when:

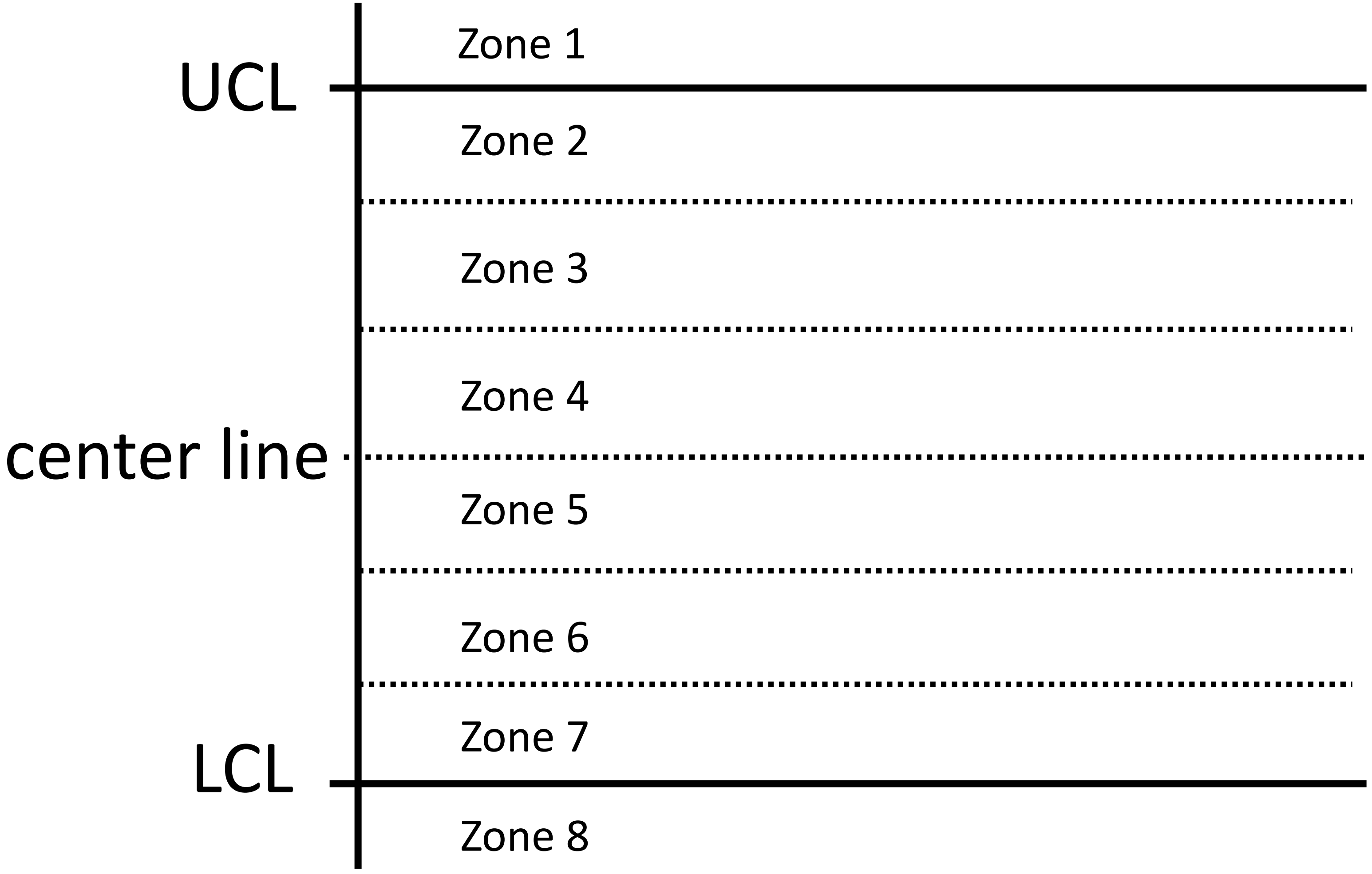
n is larger

The actual mean μ is further away from the ideal mean μ_0

<u>Our rule</u>: an alarm is defined to be an observation in Zone 1 or 8.



Be aware that more complicated rules exist.



Advantage:
More likely to detect an out of control process.

Disadvantage:
More likely to false alarm.

Be aware that many different control charts exist. Some charts, for example the Cumulative Sum (CUSUM) chart, use definitions of alarms that are not only a function of the latest collected sample, but also of previous samples.