Statistical Testing I

Motivation for Inference Testing

> **inference** - a conclusion reached on the basis of evidence and reasoning

- So, you have a bunch of data . . . Now What?
- How can I make inferences on the data?
  - Are my store sales distributed Normally?
  - Does the transit time from my supplier reduce if I switch ocean carriers?
  - Does my new order management system improve the efficiency of my warehouse?
  - What are the upper and lower bounds on the demand for my new products at, say, a 95% level of confidence?
- We will focus on three types of questions
  - Point Estimate
    - What is the average weekly sales revenue for Zippy Bright?
  - Confidence Interval Estimate
    - What are the upper and lower bounds for the average weekly sales revenue for Zippy Bright with 90% confidence?
  - Hypothesis Testing
    - Do stores A and B have the same average weekly sales revenue?
    - Is weekly demand for XP219 Normally distributed?

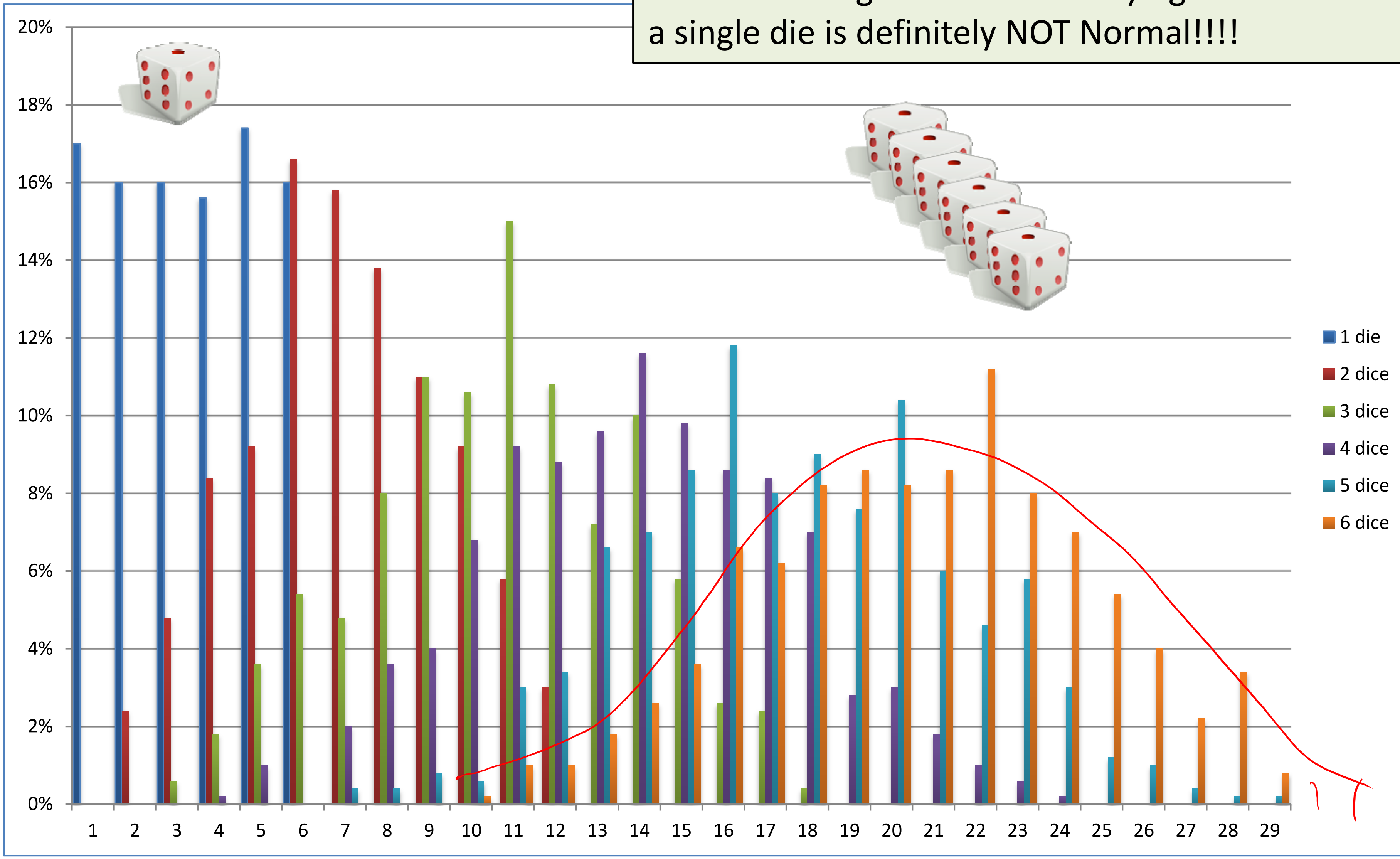
So what will we cover in this lesson?

- Some basic tools and concepts
  - Central Limit Theorem
    - Why we can use the Normal Distribution so much.
  - Sampling
    - How my sample size influences my level of confidence.
  - Confidence Intervals
    - How to establish bounds on a distribution
  - Hypothesis Testing
    - How to evaluate the outcome of a statistical test

Central Limit Theorem

Tumbling Dice . . . .

Note: As I roll more times (or use more dice) the distribution of the sum starts looking “normal-ish”! This is amazing since the underlying distribution of a single die is definitely NOT Normal!!!!



Central Limit Theorem

Recall that iid = **I**ndependent and **I**dentically **D**istributed. It means that all random variables in a distribution are independent of each other and that they are all part of the same underlying probability distribution. Think rolls of a die.

- Suppose that
  - $X_i,..X_n$ are iid with mean=μ and standard deviation = σ
  - The sum of the n random variables is S_n
  - The mean of the n random variables is $\overline{X}$

$$S_n = \sum_{i=1}^{N} X_i$$

$$\overline{X} = \frac{S_n}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

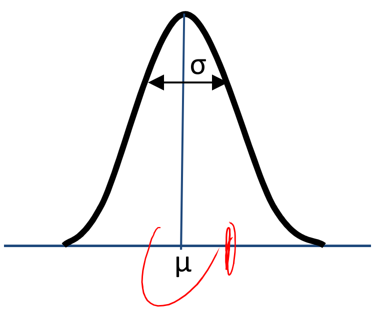
- Then, if **n** is "large"(say > 30)
  - S_n is Normally distributed with mean = **nμ** and standard deviation **$\sigma\sqrt{n}$**
  - $\overline{X}$ is Normally distributed with mean = μ and standard deviation **$\sigma/\sqrt{n}$**
- Why is this so important?
  - It does not matter what distributions the random variable X follows!
  - The distribution of the sum does not reflect the distribution of its terms
  - We will use this in forming confidence intervals and conducting tests
  - This is why we often can use the Normal distribution in practice!

Quick Review of Normal Distribution I

Quick Normal Review

Normal Distribution – symmetric around mean, μ, with a standard deviation of σ. ~N(μ, σ)

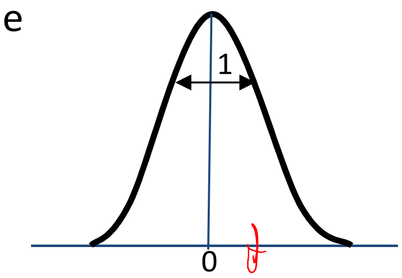
X ~N(μ, σ)



We can convert any Normal Distribution ~N(μ,σ) into the Standard Unit Normal ~N(0,1) by using the z-statistic.

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

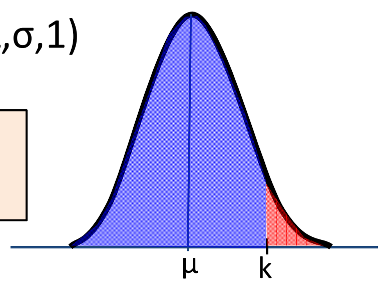
Z ~N(0, 1)



1a. What is the probability that random variable X is less than some value k?

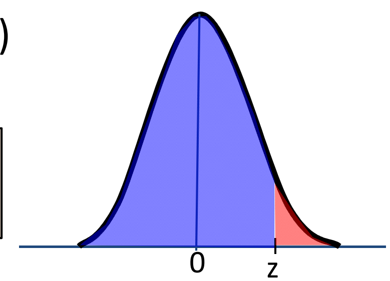
P(X≤k) =NORM.DIST(k,μ,σ,1)

Given ~N(24,8), what is P(X≤30)?
=NORM.DIST(30,24,8,1) = 0.773



P(Z≤z) =NORM.S.DIST(z,1)
z=(k-μ)/σ

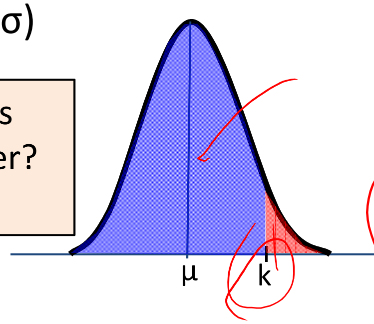
z=(30-24)/8= 0.75
=NORM.S.DIST(.75,1) = 0.773



1b. What value k will give me the probability p that random variable X is less than it?

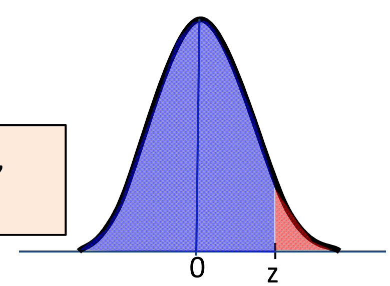
p=P(X≤k) =NORM.INV(p,μ,σ)

Given ~N(24,8) what value k gives me 90% probability that X is lower?
=NORM.INV(.90,24,8) = 34.25



P(Z≤z) =NORM.S.INV(p)
z=(k-μ)/σ

=NORM.S.INV(.90) = 1.28 so that,
k=μ+zσ = 24 + (1.28)8 = 34.25



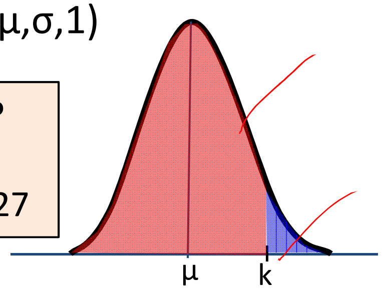
Quick Normal Review

Normal Distribution – symmetric around mean, μ, with a standard deviation of σ. ~N(μ, σ)

2a. What is the probability p that random variable X is greater than some value k?

P(X>k) =1-NORM.DIST(k,μ,σ,1)

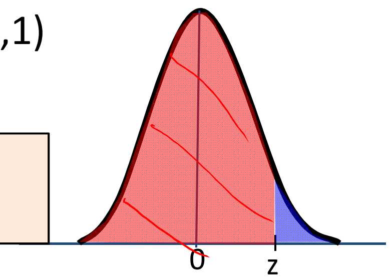
Given ~N(24,8) what is P(X>30)?
= 1 - P(X≤30)
=1-NORM.DIST(30,24,8,1) = 0.227



P(Z≤z) =1-NORM.S.DIST(z,1)

z=(k-μ)/σ

z=(30-24)/8= 0.75
=1-NORM.S.DIST(0.75,1) = 0.227



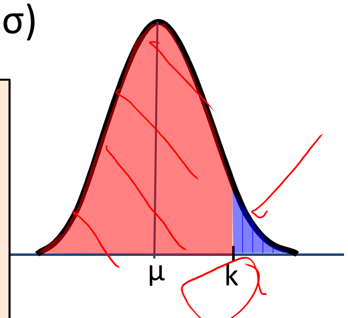
2b. What value k will give me the probability p that random variable X is greater than it?

p=P(X>k) =NORM.INV(1-p,μ,σ)

Given ~N(24,8) what k gives me 10% probability that X is greater?

Note that this k is the same point where P(X≤k)= 1 – 0.10 = 0.90!

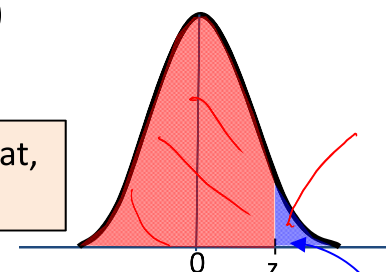
=NORM.INV((1-0.10),24,8) = 34.25



P(Z>z) =NORM.S.INV(1-p)

z=(k-μ)/σ

=NORM.S.INV(1-0.10) = 1.28 so that,
k=μ+zσ = 24 + (1.28)8 = 34.25



If we call this area α
z=NORM.S.INV(1-α)

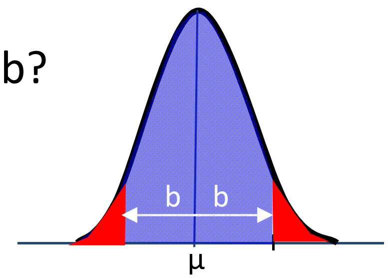
MIT Center for Transportation & Logistics

Quick Review of Normal Distribution II

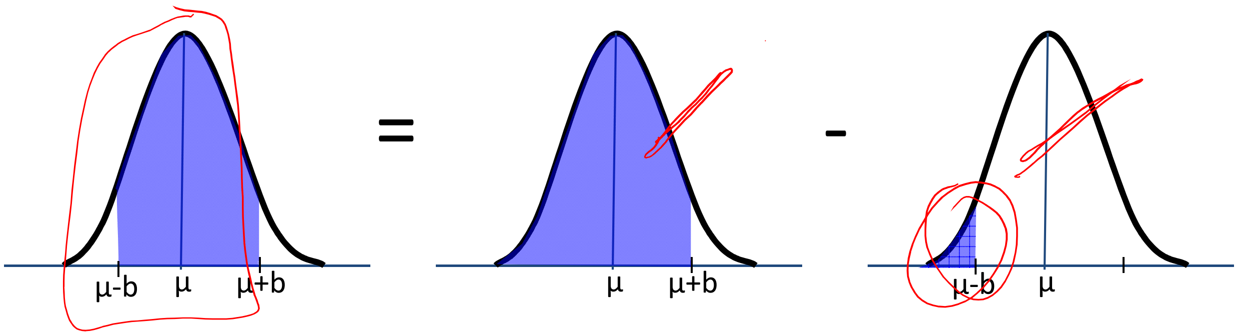
Quick Normal Review

Normal Distribution – symmetric around mean, μ, with a standard deviation of σ. ~N(μ, σ)

3a. What is the probability that random variable X is between μ-b and μ+b?

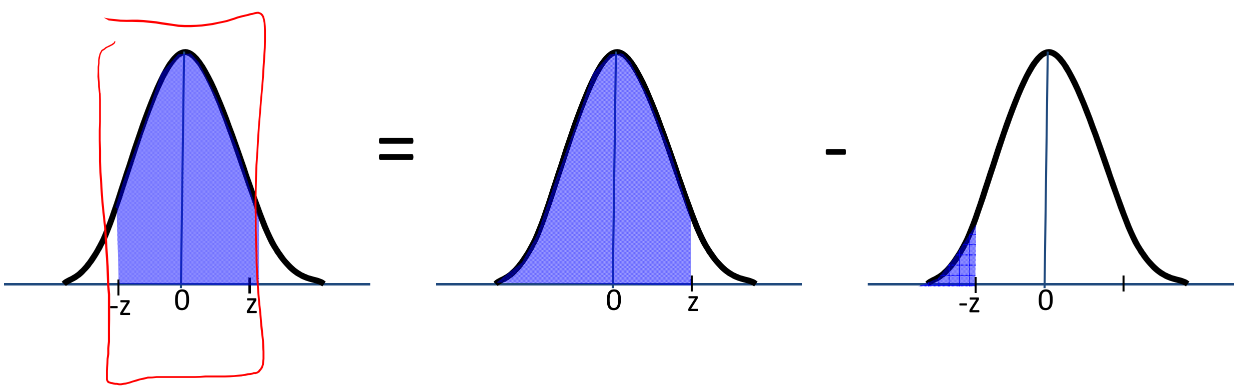


P(μ-b≤X≤μ+b) = NORM.DIST(μ+b,μ,σ,1) - NORM.DIST(μ-b,μ,σ,1)



Given ~N(24,8) what is
P(20≤X≤28), that is, b=4?
= NORM.DIST(28,24,8,1)
- NORM.DIST(20,24,8,1)
= 0.69 – 0.31 = 0.38

P(-z≤Z≤z) = NORM.S.DIST(z,1) - NORM.S.DIST(-z,1)



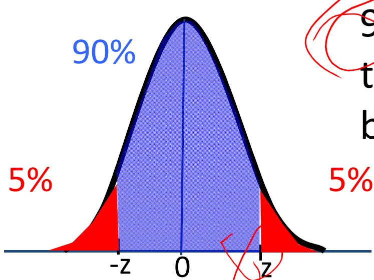
z=(28-24)/8= 0.5
-z=(20-24)/8 = -0.5
= NORM.S.DIST(0.5,1)
- NORM.S.DIST(-0.5,1)
= 0.69 – 0.31 = 0.38

Quick Normal Review

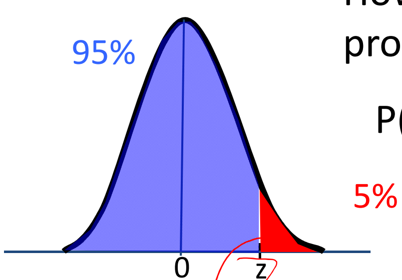
Normal Distribution – symmetric around mean, μ, with a standard deviation of σ. ~N(μ, σ)

3b. What bounds b are required so that the probability that random variable X is between μ-b and μ+b is equal to p?

Given ~N(24,8) find the values for b such that the P(24-b≤X≤24+b) = 0.90
Find bounds (μ +/- b) such that we have a 90% probability of falling within them.
For Standard Normal, find z such that P(-z≤Z≤z) = 0.90



90% within the bounds means that 10% of the probability distribution is outside the bounds; 5% on each side of the tail!



How do I find the z value where the probability of the right tail is 5%?

P(Z≤z) = NORM.S.INV(1-.05) = 1.64

To find the z value where α is the probability outside the bounds, simply find P(Z≤z) = 1 – α/2

What are the bounds for 90%?
α=0.10 so 1 – α/2 = 0.95
= NORM.S.INV(0.95) =1.64
So, b = (1.64)8= 13.12
Upper bound = 24 + 13.12 = 37.12
Lower bound = 24 - 13.12 = 10.88

There is a **90%** probability that the random variable, X, will fall between **10.88** and **37.12**!

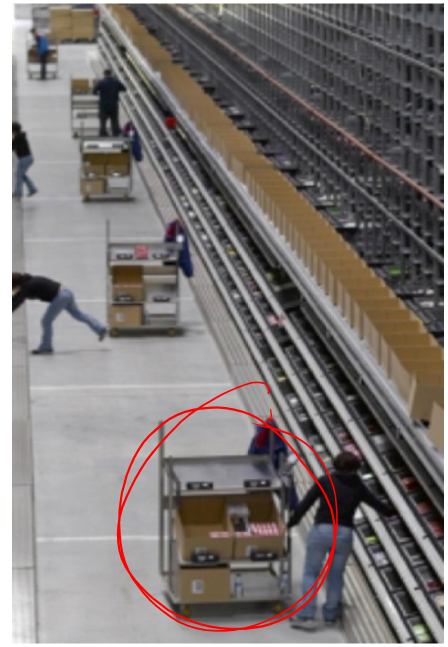
What are the bounds for 95%?
α=0.05 so 1 – α/2 = 0.975
= NORM.S.INV(0.975) = 1.96
So, b = (1.96)8= 15.68
Upper bound = 24 + 15.68 = 39.68
Lower bound = 24 – 15.68 = 8.32

There is a **95%** probability that the random variable, X, will fall between **8.32** and **39.68**!

Applying Central Limit Theorem

Example: Order Picking I

- You are managing an e-commerce fulfillment center where a team of pickers assemble an order to be ready for shipping. On average, your teams can pick and assemble 4 orders with a standard deviation of 5 orders within a minute. There are always orders in the queue waiting to be picked.
- What is the approximate probability that they will pick at least 260 orders within an hour?

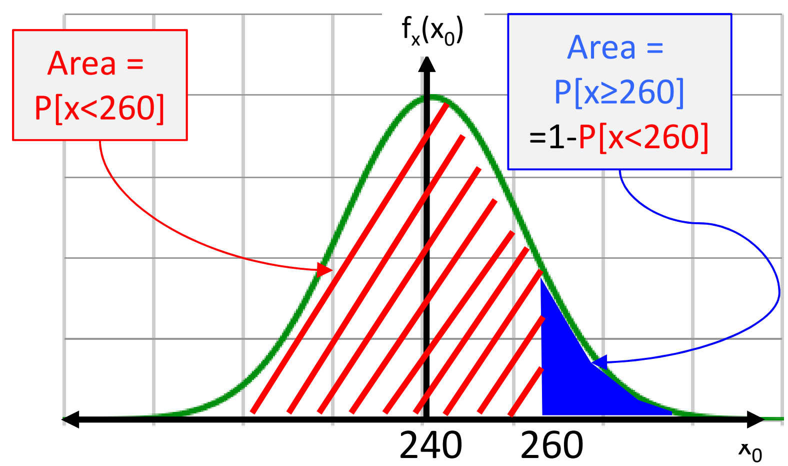


What do we know?

- μ= 4 orders and σ=5 orders per minute
- Since there are 60 minutes in an hour:
  - $E[S_{60}] = 60(4) = 240$ orders
  - $StdDev[S_{60}] = \sqrt{60}(5) = 38.7$ orders

The Central Limit Theorem allows us to assume that we will approximate the Normal distribution!

- Total Orders Picked in an Hour ~ N(240, 38.7)
- So, we want to find P[X≥260]

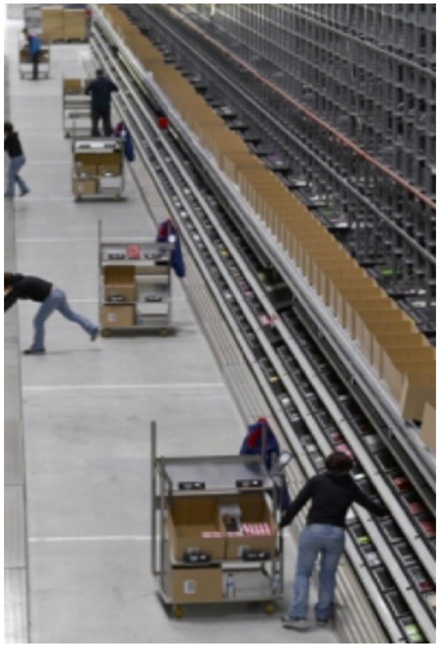


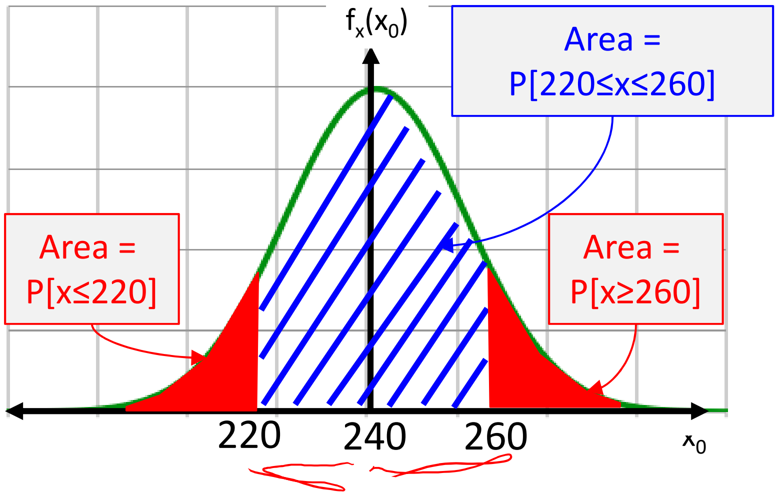
P[Orders≥260] = 1 – P[Orders<260]
= 1 - NORM.DIST(260,240, 38.7, 1)
= 1 – 0.697 = 0.303 or **~30%**

Answer: About 30% of the time they will pick at least 260 orders in an hour.

Example: Order Picking II

- Same situation where expected orders picked in an hour are distributed Normally with mean of 240 and standard deviation of 38.7.
- What is the approximate probability that they will pick between 220 and 260 orders within an hour?





I. Find and use z values
We know that z=(x-μ)/σ, so
 z=(260-240)/38.7 = 0.517 for upper bound
 z=(220-240)/38.7 = -0.517 for lower bound
We want P[-.517≤z≤0.517] =P[z<0.517] – P[z<-0.517]
 = NORM.S.DIST[0.517,1] – NORM.S.DIST[-0.517,1]
Which gives us 0.697 – 0.303 = 0.394 ~ **40%**

II. Look at Symmetry
We know that P[x≥260] = 30% and that Normal Distributions are symmetric. Therefore, P[x≤220] = 30% and **P[220≤x≤260] = 40%**

Answer: About 40% of the time they will pick between 220 and 260 orders in an hour.

Sampling

MoonDoe Café

- MoonDoe Café is a nationwide chain of 2600 coffee shops serving over a million customers a day. They sell a mix of hot and cold beverages as well as healthy snacks. MoonDoe believes that the more time that a customer spends within its shop, the more likely they are to spend even more money. MoonDoe has hired you to understand this relationship a little better and to see how it can be improved.
- Specifically, your first tasks are to:
  - Estimate the expected duration of customer visits to a MoonDoe café and
  - Provide some confidence intervals for your estimate.
- How should you proceed?
  - Can you ask everyone?
  - No! So you need to sample!
  - What could possibly go wrong?

Statistical Sampling

- Terms
  - **Population** – e.g., all 1 million daily customers that visit MoonDoe cafes
    - Universal set of all items of interest
    - Parameters used to describe the distribution of random variable X (e.g., $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$)
  - **Sample** – a smaller set of customer visits, say, 50
    - Sub set of the population items
    - Distinguish between before and after sample is taken
    - Statistics are used to describe the observed values x (e.g., $\bar{\mathbf{x}}$ and **s**)
    - Sample mean ($\bar{X}$) and Sample standard deviation (S) are random variables
  - **Random Sample**
    - Sample selected from population so that each item is equally likely
- Objective
  - Estimate the <u>true population</u> parameters by using the <u>sample</u> parameters:
    - True mean ($\boldsymbol{\mu}$) of random variable X estimated by using the observed sample mean $\bar{\mathbf{x}}$
    - True standard deviation ($\boldsymbol{\sigma}$) of random variable X estimated by using the observed sample standard deviation **s**

MoonDoe Café

Why 50? (we'll discuss sample size a little bit later)

- You have collected a random sample of the duration of 50 MoonDoe customer visits.
- You can easily calculate:
  - Mean= $\bar{x}$ = 24.6 minutes and Standard Deviation = s = 10.7
- Are we done?

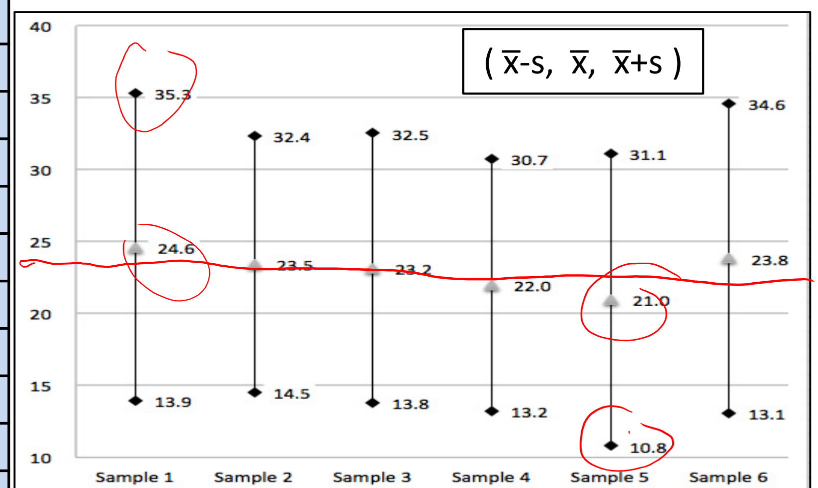
Sample 1

6.7	23.6	12.3
37.2	13.6	26.3
20.7	39.7	29.5
23.9	21.9	29.8
34.6	16.0	7.3
52.3	16.9	23.2
28.9	29.1	18.6
42.5	18.8	28.5
32.3	22.5	33.8
36.7	39.7	13.5
32.8	6.0	14.9
24.4	33.1	6.1
27.1	6.5	33.3
19.6	30.8	31.4
28.3	20.2	13.4
22.6	16.0	11.0
34.8	38.3	

Sample 2
$\bar{x}$ = 23.5 s=8.9

15.5	24.3	31.1
20.6	23.9	19.4
21.4	34.6	17.2
21.8	21.5	15.2
27.1	32.6	21.0
16.0	29.9	14.7
38.9	18.1	24.5
14.8	6.1	41.4
17.2	4.0	33.3
13.3	18.6	34.5
41.3	28.6	28.8
29.6	15.7	18.6
9.8	22.3	31.3
24.1	11.2	31.8
28.6	36.0	18.9
19.6	19.3	34.6
16.0	34.3	

Let's plot the mean and the plus/minus a standard deviation for six samples . . .



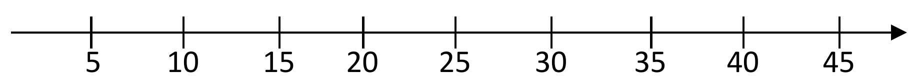
So, how confident am I that the true population mean is within some range or interval of my sample mean?

Confidence Intervals I

MoonDoe Café – Confidence Intervals

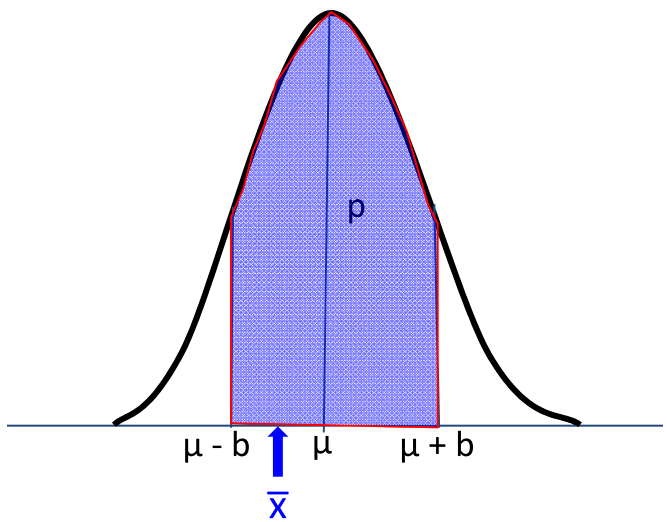
- Let's create a 90% confidence interval for the true population mean of the duration of MoonDoe customer visits.
- What does this mean?
  - We want to find bounds from the sample mean such that the population mean will fall between them 90% of the time.
- Some things to think about
  - How will the size of my sample, n, impact the confidence interval?
  - How will the sample standard deviation, s, impact the confidence interval?
  - How will the sample mean, $\bar{x}$, impact the confidence interval?
  - How will the probability selected impact the confidence interval?
  - What are the random variables here?
    - The true population mean, μ, is NOT a random variable
    - The sample mean, $\bar{X}$, IS a random variable where $\bar{x}$ is an observed sample mean.

MoonDoe customer visit duration in minutes



Confidence Intervals

The underlying true population distribution with a mean μ and standard deviation σ.



Recall from the Central Limit Theorem:

- The expected value of the sample mean is the population mean and
- The standard deviation of the sample mean is the population mean divided by the sample size, n.

$$E\left[\bar{X}\right] = \mu \qquad \sigma_{\bar{X}} = \sigma / \sqrt{n}$$

This says that the probability that the random variable, $\bar{X}$, falls within b units plus or minus of the population mean, μ, is p.

$$p = P\left(\mu - b \le \bar{X} \le \mu + b\right) = 0.90$$

Simplify this by subtracting μ from all sides . . .

$$p = P\left(-b \le \bar{X} - \mu \le b\right) = 0.90$$

Divide the terms by σ/√n . . .

$$p = P\left(-\frac{b}{\sigma/\sqrt{n}} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \frac{b}{\sigma/\sqrt{n}}\right) = 0.90$$

We know that from the Central Limit Theorem that this will approximate the Normal Distribution for large samples. This should look familiar! The Z statistic – subbing in . . .

$$p = P\left(-\frac{b}{\sigma/\sqrt{n}} \le Z \le \frac{b}{\sigma/\sqrt{n}}\right) = 0.90$$

Confidence Intervals

This is where we last were . . .

$$p = P\left(-\frac{b}{\sigma/\sqrt{n}} \le Z \le \frac{b}{\sigma/\sqrt{n}}\right) = 0.90$$

Earlier, we saw that the z for P(-z≤Z≤z) = 0.90 is the same z value such that P(Z≤z) = 0.95, which is z=1.64.

So, if z = 1.64 = b/(σ/√n), then **b= 1.64 σ/√n**.

Substituting for b in our original probability equation:

$$p = P\left(\mu - \frac{1.64\sigma}{\sqrt{n}} \le \bar{X} \le \mu + \frac{1.64\sigma}{\sqrt{n}}\right) = 0.90$$

Rearranging the terms gives us:

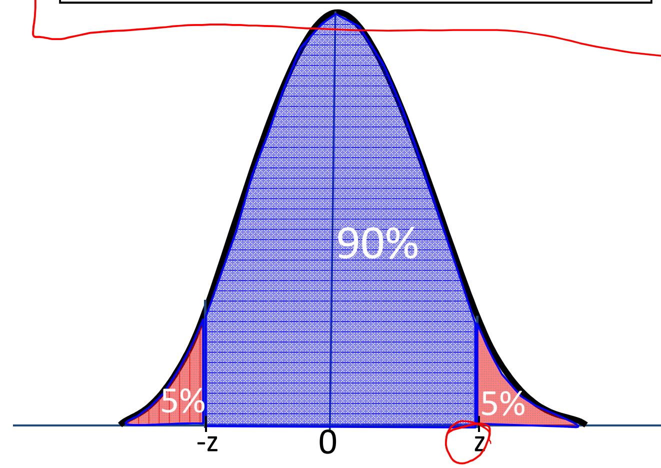
$$p = P\left(\bar{X} - \frac{1.64\sigma}{\sqrt{n}} \le \mu \le \bar{X} + \frac{1.64\sigma}{\sqrt{n}}\right) = 0.90$$

Giving us the 90% Confidence Interval for random variable X for large n (>30):

$$\left[\bar{x} - \frac{1.64s}{\sqrt{n}},\ \bar{x} + \frac{1.64s}{\sqrt{n}}\right]$$

This is where we started. . .

$$p = P\left(\mu - b \le \bar{X} \le \mu + b\right) = 0.90$$



a very quick aside . . .

We start with this: $\mu - \frac{1.64\sigma}{\sqrt{n}} \le \bar{X} \le \mu + \frac{1.64\sigma}{\sqrt{n}}$

Subtract μ from all sides: $-\frac{1.64\sigma}{\sqrt{n}} \le \bar{X} - \mu \le \frac{1.64\sigma}{\sqrt{n}}$

Subtract $\bar{X}$ from all sides: $-\frac{1.64\sigma}{\sqrt{n}} - \bar{X} \le -\mu \le \frac{1.64\sigma}{\sqrt{n}} - \bar{X}$

Multiply all sides by -1 (reversing signs!): $\frac{1.64\sigma}{\sqrt{n}} + \bar{X} \ge \mu \ge -\frac{1.64\sigma}{\sqrt{n}} + \bar{X}$

Rearrange: $\bar{X} - \frac{1.64\sigma}{\sqrt{n}} \le \mu \le \bar{X} + \frac{1.64\sigma}{\sqrt{n}}$

Note: We use the observed sample mean, x̄, and standard deviation, s, to estimate the bounds since we do not know the population values!

MIT Center for Transportation & Logistics

Finding Confidence Intervals for MoonDoe I

MoonDoe Café – Confidence Intervals

- Let's create a 90% confidence interval for the true population mean of the duration of MoonDoe customers.
- You have collected a random sample of 50 customer visits.
  - sample size = n = 50
  - sample mean = $\bar{x}$ = 24.6 minutes
  - sample standard deviation = s = 10.7
  - we know the confidence value = c = 1.64
- So, the 90% Confidence Interval is [22.1, 27.1]

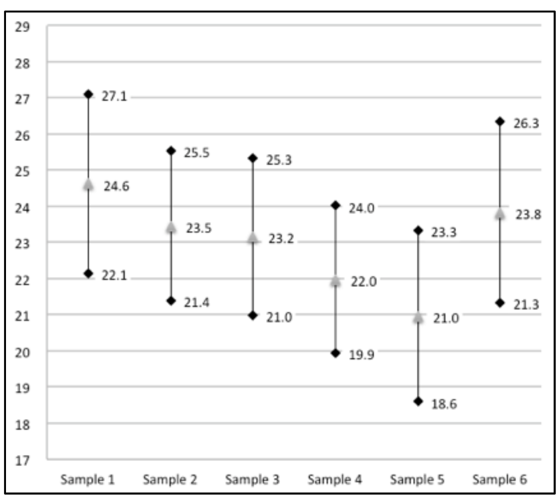
$$\left[\bar{x} - \frac{1.64s}{\sqrt{n}},\ \bar{x} + \frac{1.64s}{\sqrt{n}}\right]$$

$$\left[24.6 - \frac{1.64(10.7)}{\sqrt{50}},\ 24.6 + \frac{1.64(10.7)}{\sqrt{50}}\right]$$

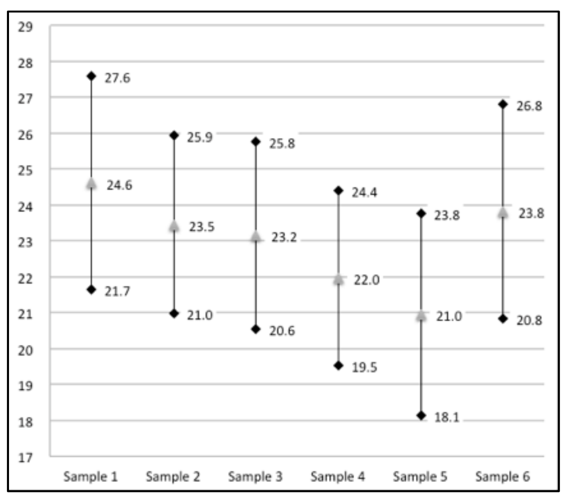
$$\left[24.6 - 2.48,\ 24.6 + 2.48\right]$$

$$= \left[22.12,\ 27.08\right]$$

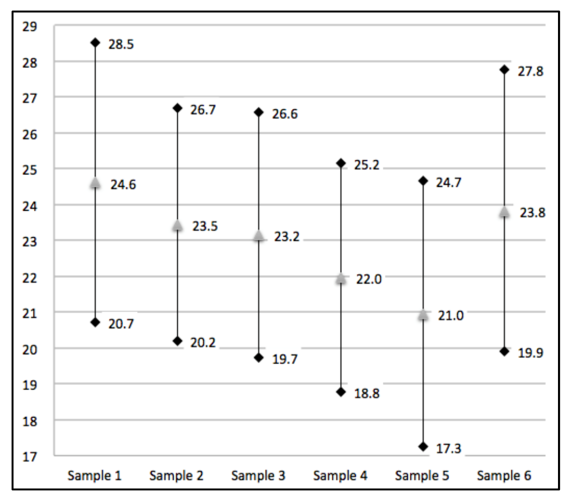
90% Confidence Intervals



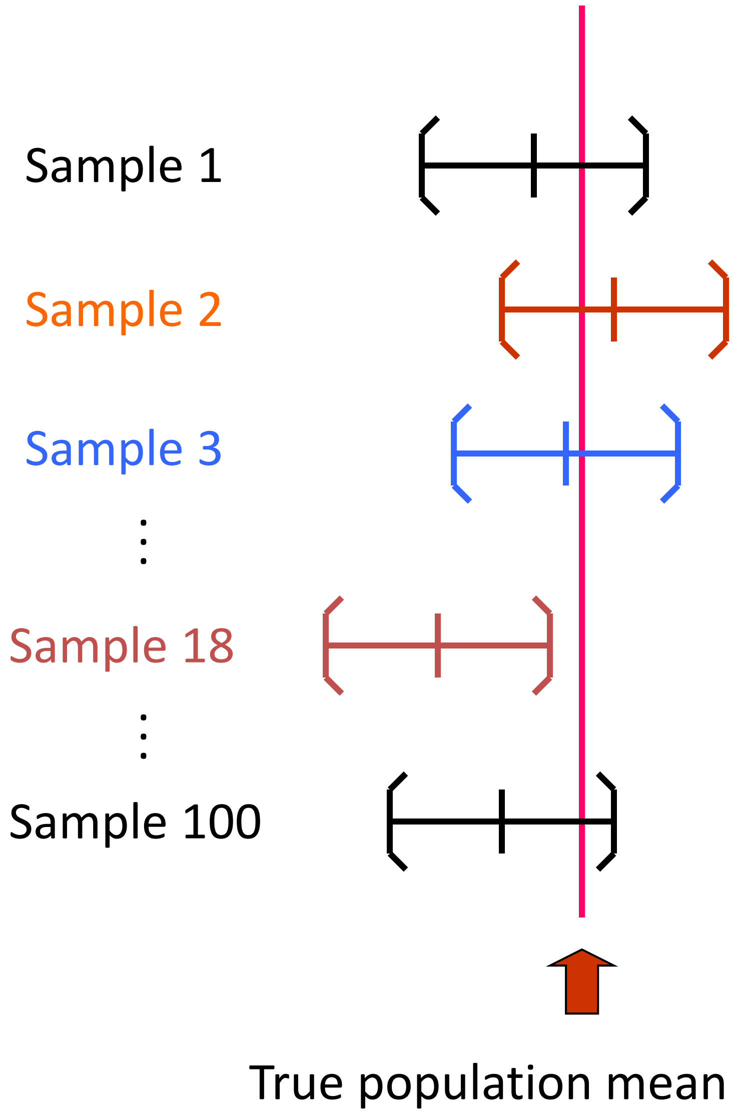
95% Confidence Intervals



99% Confidence Intervals



Interpreting Confidence Intervals



In a typical application, we only sample once and report a single confidence level, for example, 95%.

If we repeated this sampling procedure 100 times, our (random) intervals will capture the true population mean, on average, 95 times out of the 100.

Confidence Intervals II
smaller sample size

MoonDoe Café

Why 15? (it was all you were able to collect)

- You have collected a random sample of the duration of 15 MoonDoe customer visits.
  - sample size = n = 15
  - sample mean = $\bar{x}$ = 19.6 minutes
  - sample standard deviation = s = 10.5 minutes

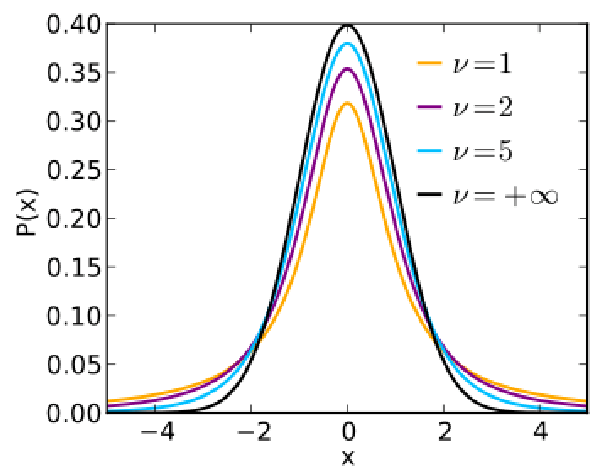
1	14.8
2	47.0
3	17.9
4	18.0
5	15.5
6	16.3
7	6.4
8	10.2
9	27.4
10	22.8
11	34.5
12	12.9
13	17.0
14	23.2
15	10.1

Can we use our earlier results where the confidence value, c, is found with Normal Distribution ?

$$\left[\bar{x}-\frac{cs}{\sqrt{n}},\ \bar{x}+\frac{cs}{\sqrt{n}}\right]$$

Almost. We found c using the Normal Distribution due to the Central Limit Theorem. But, this requires a "large" sample size (generally n>30).

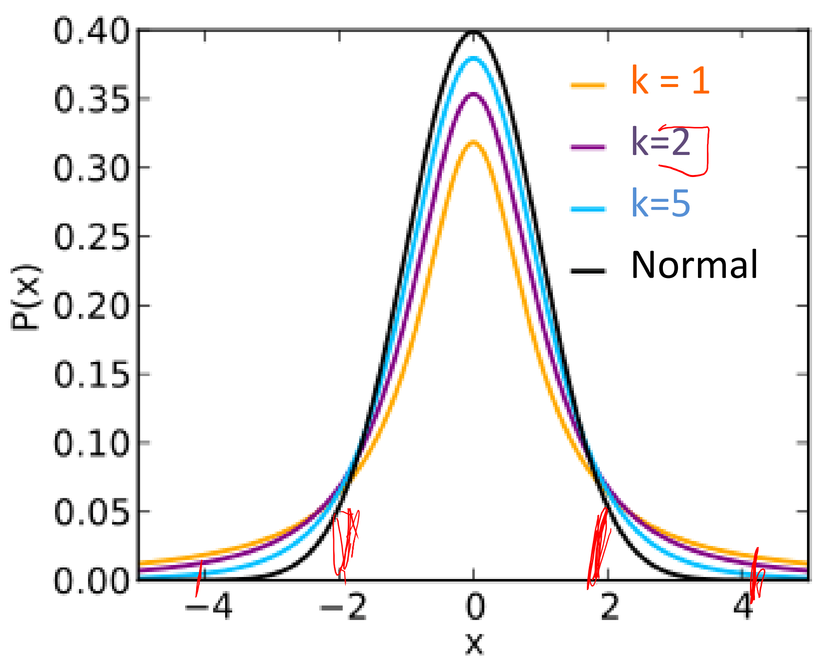
We need to look at a different distribution to calculate the confidence value, c, specifically for small sample sizes.
It is called the t-distribution.



By Skbkekas - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9546828

Student t-Distribution

- The t-distribution . . .
  - is bell-shaped and symmetric around 0,
  - is flatter and has fatter tails than the Normal distribution,
  - its shape is a function of k, the degrees of freedom, where k=n-1, and
  - its mean=0 and its standard deviation = √(k/k-2).



For the Normal distribution, 95% falls within +/- z=1.95 while for the t-distribution (k=2) it is +/- 4.30!

But, as n gets bigger the t-distribution closely approximates the Normal. When n>30 they are essentially identical.

We will use the t-distribution to find c for small sample sizes just like we used the Normal distribution for large.

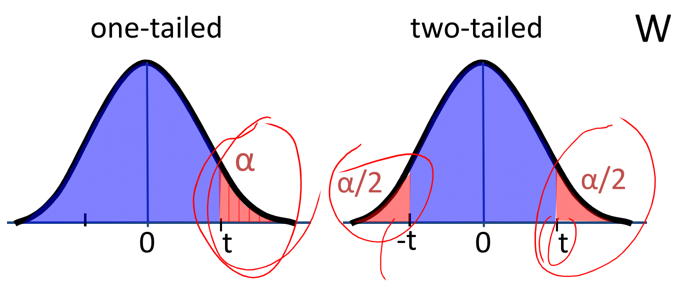
We need to remember whether we should be using a one or two tailed test:

1. For a random variable to be greater than or less than some value (one-tailed)
2. For a random variable to be within some confidence intervals (two-tailed)

t-Distribution Functions for Spreadsheets

warning! there are other functions out there depending on the software version

Function	Microsoft Excel	Google Sheets	LibreOffice->Calc
cdf of t Distribution (returns α probability)	=T.DIST.2T(X, k) =T.DIST(X, k) =TDIST(X, k, num_tails)	=TDIST (X, k, num_tails)	=T.DIST.2T(X, k) =T.DIST(X, k) =TDIST(X, k, num_tails)
Inverse of t Distribution (returns the t-value)	=T.INV.2T(α, k)	=T.INV.2T(α, k)	=T.INV.2T(α, k)



Where

k= degrees of freedom (generally n-1)

α = alpha probability, e.g., if 95% Confidence Interval then α=0.05

Note that we can convert the two tail into a one-tail test using 2α

num_tails = 1 for one tailed test, 2 for two tailed test

Examples :

1. Find t-value for 90% Confidence Interval with 17 observations
   - k=n-1 = 17-1 = 16 α=1-0.90 = 0.10 this is two-tailed
   - =T.INV.2T(.10,16) = 1.75
2. What is the probability that my t value is greater than 1.95 with 8 observations?
   - k=n-1 = 8-1 = 7 X=1.95 this is one-tailed
   - =TDIST(1.95,7,1) = 0.046 or 4.6%
3. What Confidence Interval is provided with a t-value of 2.65 with 21 observations?
   - k=n-1 = 21-1 = 20 X=2.65 this is two-tailed
   - =TDIST(2.65,20,2) = 0.015 so, 1 – 0.015 = 0.985 or 98.5% Confidence Interval
   - Or, we could have used =T.DIST.2T(2.65,20) = 0.015 to get same answer.

MoonDoe Café

- You have collected a random sample of the duration of 15 MoonDoe customer visits. Let's create a 90% confidence interval for the true population mean of the duration of MoonDoe customers using this small sample.
  - sample size = n = 15
  - sample mean = $\bar{x}$ = 19.6 minutes
  - sample standard deviation = s = 10.5 minutes

$$\left[\bar{x} - \frac{cs}{\sqrt{n}},\ \bar{x} + \frac{cs}{\sqrt{n}}\right]$$

1	14.8
2	47.0
3	17.9
4	18.0
5	15.5
6	16.3
7	6.4
8	10.2
9	27.4
10	22.8
11	34.5
12	12.9
13	17.0
14	23.2
15	10.1

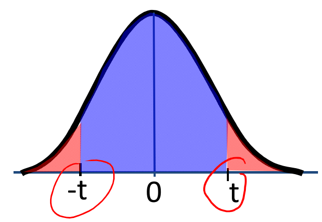
Find the value for c for 90% Confidence Interval:

k = n - 1 = 15 - 1 = 14

α = 1 - 0.90 = 0.10

this is two-tailed

c = T.INV.2T(.10,14) = 1.76



$$\left[19.6 - \frac{1.76(10.5)}{\sqrt{15}},\ 19.6 + \frac{1.76(10.5)}{\sqrt{15}}\right] = \left[19.6 - 4.77,\ 19.6 + 4.77\right] = \left[14.83,\ 24.37\right]$$

The 90% Confidence Interval based on this sample is [14.8, 24.4]

Note:

- As expected, this CI is larger than the CIs found using bigger samples.
- Also, the 95% CI would be even larger - *try it!* I got [13.8, 25.4]
- The only difference in CIs for Normal and t distributions is how we find c!

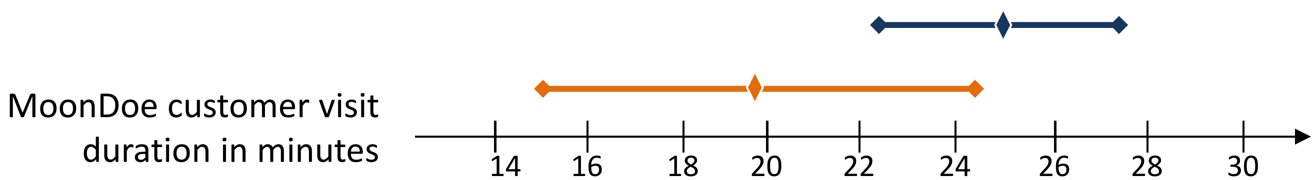
Confidence Intervals III

Comparing Confidence Intervals

- We created two 90% confidence intervals for the true population mean of the duration of MoonDoe customer visits.

Small Sample (15 visits)
- $\bar{x}$ = 19.6 minutes
- s = 10.5 minutes
- 90% Confidence Interval = [14.8, 24.4]

Large Sample (50 visits)
- $\bar{x}$ = 24.6 minutes
- s = 10.7 minutes
- 90% Confidence Interval = [22.1, 27.1]



How should we interpret Confidence Intervals?

- If we repeat this sampling procedure an infinite number of times, the true mean of this population should appear in 90% of those confidence intervals
- This does <u>not</u> mean that 90% of observations or data points from this population will fall within these confidence intervals.

Confidence vs. Prediction Intervals

Large Sample (50 visits)

$\bar{x}$ = 24.6 minutes

s = 10.7 minutes

c = 1.64 = NORM.S.DIST(0.95)

Confidence Interval

$$\left[\bar{x} - \frac{cs}{\sqrt{n}},\ \bar{x} + \frac{cs}{\sqrt{n}}\right]$$

If we repeat this sampling an infinite number of times, the true mean of this population should appear in 90% of those confidence intervals

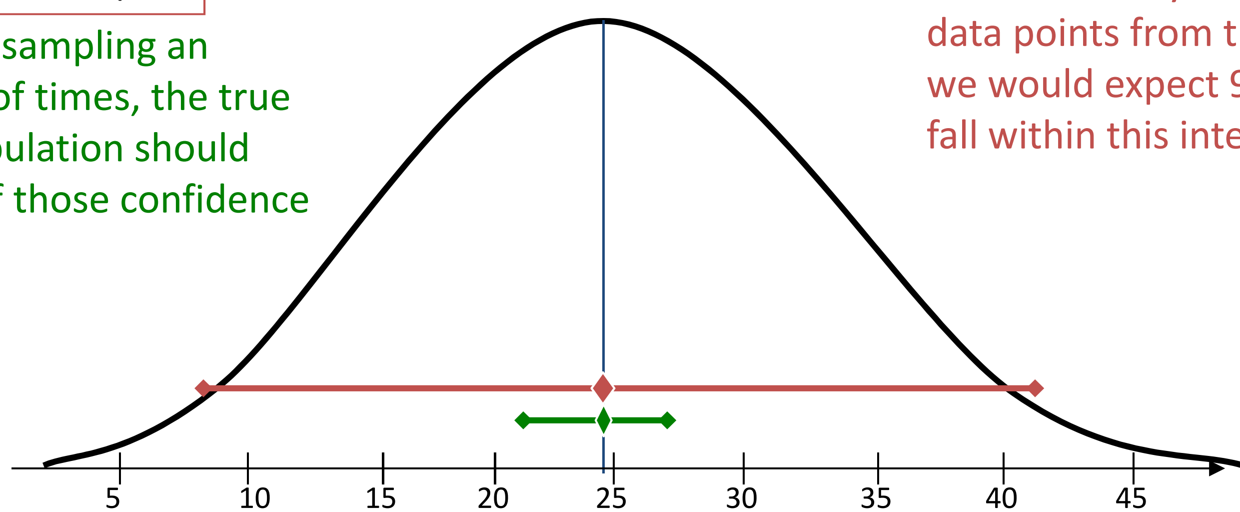
24.6 +/- 2.5
[22.1, 27.1]

Prediction Interval

$$\left[\bar{x} - cs,\ \bar{x} + cs\right]$$

If we randomly choose other data points from this population, we would expect 90% of them to fall within this interval.

24.6 +/- 17.5
[7.1, 42.1]



How would your Confidence and Prediction Intervals change, if you sampled 500 visits instead of 50 assuming your $\bar{x}$ and s did not change?

24.6 +/- (1.64)10.7/√500
24.6 +/- 0.8
[23.8, 24.4]

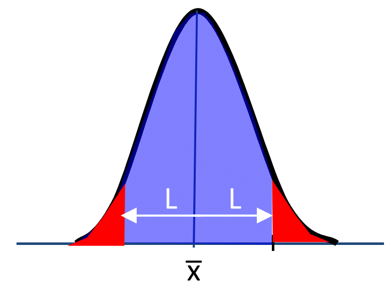
No Change!
24.6 +/- 17.5
[7.1, 42.1]

MIT Center for Transportation & Logistics

Insights on Confidence Intervals for Mean

$$\left[\bar{x} - \frac{cs}{\sqrt{n}}, \bar{x} + \frac{cs}{\sqrt{n}}\right]$$

- c = Confidence Value – distribution based on sample size:
  - <30 t-distribution
  - ≥30 Normal distribution
- L = interval = cs/√n



- Tradeoffs between interval, sample size, and confidence.
  - If <u>**n** is fixed</u>, using a higher confidence value **c** leads to a wider interval **L**.
  - If <u>**c** is fixed</u>, increasing sample size **n** leads to smaller interval **L**.
  - If <u>both **n** & **c** are fixed</u>, reducing variability **s** leads to smaller interval **L.**

- How big should my sample be?
  - Obviously, bigger is better, but how big is good enough?
  - We can set n to meet a specific tolerance level L and c.
    - Need to run a pilot or estimate sample standard deviation s to use.

$$L = \frac{cs}{\sqrt{n}} \qquad \boxed{n = \frac{c^2 s^2}{L^2}}$$

Example: How big should my MoonDoe sample be so that I am within +/- 2 minutes of the population mean at a 99% Confidence Level?

- We estimated s (take from 15 visit sample) = 11
- Since L=2, c=NORM.S.INV(.995) = 2.56

$$n = \frac{(2.56)^2 (11)^2}{(2)^2} = \boxed{198}$$

Hypothesis Testing Basics

Hypothesis Testing

- Method for making a choice between two **mutually exclusive** and **collectively exhaustive** alternatives
- We make two hypotheses:
  - Null Hypothesis (H_0)
    - Cannot be rejected unless data argues overwhelmingly otherwise
  - Alternative Hypothesis (H_1)
    - The other possible outcomes – mutually exclusive of H_0
- We test, <u>at a specified significance level</u>, to see if we can:
  - Reject the Null Hypothesis, or
  - Fail to reject the Null Hypothesis

Note: Technically, we can never actually "accept" the Null Hypothesis, only choose to not reject it.

Some examples for the MoonDoe Café problem where we are comparing visit duration (D_i) between store i and the rest.

H_0: $D_i \geq D_{MD}$
H_1: $D_i < D_{MD}$

H_0: Visit durations at store i are not shorter than at others at significance level α

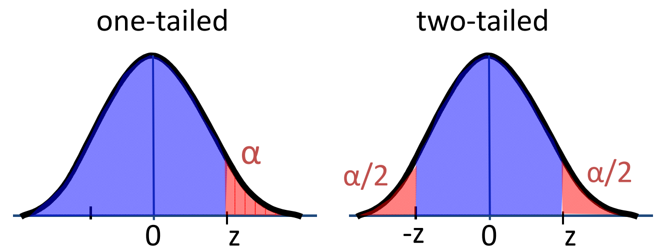
H_0: $D_i = D_{MD}$
H_1: $D_i \neq D_{MD}$

H_0: Visit durations at store i are the same as those at others at significance level α

H_0: $D_i \leq D_{MD}$
H_1: $D_i > D_{MD}$

H_0: Visit durations at store i are not longer than at others at significance level α

Hypothesis Testing



1. Select the test statistic of interest
   - What are you measuring?
   - Sample size, n, dictates distribution to use (Normal or t).
2. Determine whether this is a one or two tailed test
   - If measuring "any difference" or "is not the same as", then two tailed.
   - If measuring "is greater than" or "is less than", then use one tailed.
3. Pick your significance level and critical value
   - Set alpha, typical values 0.05 or 0.01.
   - Find corresponding z or t-statistic
4. Formulate your Null & Alternative hypotheses
   - H_0: The distributions are the same
   - H_1: Statement for test statistic difference (greater, less, or different)
5. Calculate the test statistic
6. Compare the test statistic to the critical value
   - Reject or Do Not Reject the Null Hypothesis at selected level of significance

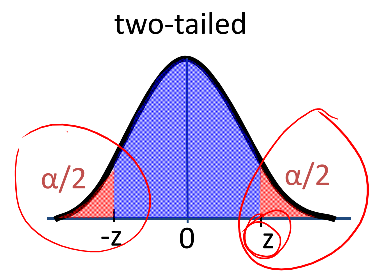
Hypothesis Testing Example

Hypothesis Testing: NextGen Stores

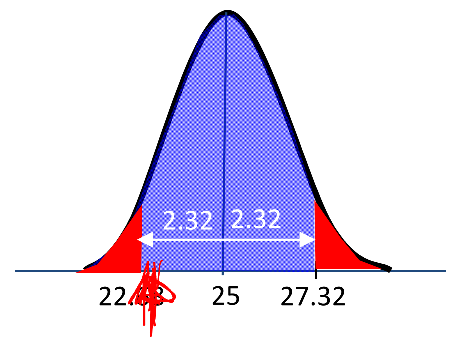
- MoonDoe Café is exploring ways to increase visit duration (and thus sell more coffee). They have piloted a next generation store that features more seating area, better WiFi connections, and special amenities. Of course, these improvements require significant investment and management wants to know if this NextGen store has actually changed the visit duration.
- From earlier studies, you have determined that the distribution for visit duration has a mean of 23 minutes. You have sampled 100 visits for the NextGen stores and they have a sample mean of 25 minutes and a sample standard deviation of 9 minutes.
- Set up a Hypothesis Test to test whether the average visit duration for the NextGen store is different from that of regular cafés.

Hypothesis Testing: NextGen Stores

1. Select the test statistic of interest
   - We will use average visit duration, $\bar{x}_{NextGen}$ and μ, in minutes as the statistic
   - Since we have 100 observations in our sample (n=100) we can use Normal
2. Determine whether this is a one or two tailed test
   - We are testing if there are any differences – longer or shorter – so it is two-tailed
3. Pick your significance level and critical value
   - Set alpha = 1% or 0.01
   - The corresponding z value = NORM.S.INV(1-α/2)=2.58



4. Formulate your Null & Alternative hypotheses
   - H_0: $\bar{x}_{NextGen} = \mu_{Current}$
   - H_1: $\bar{x}_{NextGen} \neq \mu_{Current}$
5. Calculate the test statistic
   - Is the population mean, $\mu_{Current}$, within the 99% CI of sample?
   - 99% CI = $\bar{x}_{NextGen}$ +/- cs/√n = 25 +/- (2.58)(9)/√100 = 25 +/- 2.32

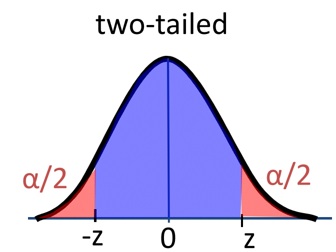


6. Compare the test statistic to the critical value
   - Population mean μ =23 and is within this 99% Confidence Interval
   - We CANNOT REJECT the null hypothesis that the average visit duration are the same at current and NextGen stores at a 1% level of significance.

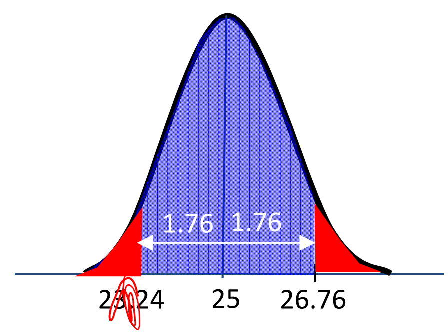
OK – but what about at the 5% level of significance?

Hypothesis Testing: NextGen Stores

1. Select the test statistic of interest
   - We will use average visit duration, $\overline{x}_{NextGen}$ and μ, in minutes as the statistic
   - Since we have 100 observations in our sample (n=100) we can use Normal
2. Determine whether this is a one or two tailed test
   - We are testing if there are any differences – longer or shorter – so it is two-tailed
3. Pick your significance level and critical value
   - Set alpha = 5% or 0.05
   - The corresponding z value = NORM.S.INV(1-α/2)=1.96



4. Formulate your Null & Alternative hypotheses
   - H_0: $\overline{x}_{NextGen} = \mu_{Current}$
   - H_1: $\overline{x}_{NextGen} \neq \mu_{Current}$
5. Calculate the test statistic
   - Is the population mean, $\mu_{Current}$, within the 99% CI of sample?
   - 95% CI = $\overline{x}_{NextGen}$ +/- cs/√n = 25 +/- (1.96)(9)/√100 = $\boxed{25 \text{ +/- } 1.76}$



6. Compare the test statistic to the critical value
   - Population mean μ =23 and is outside of this 95% Confidence Interval
   - We REJECT the null hypothesis that the average visit duration are the same at current and NextGen stores at a 5% level of significance.

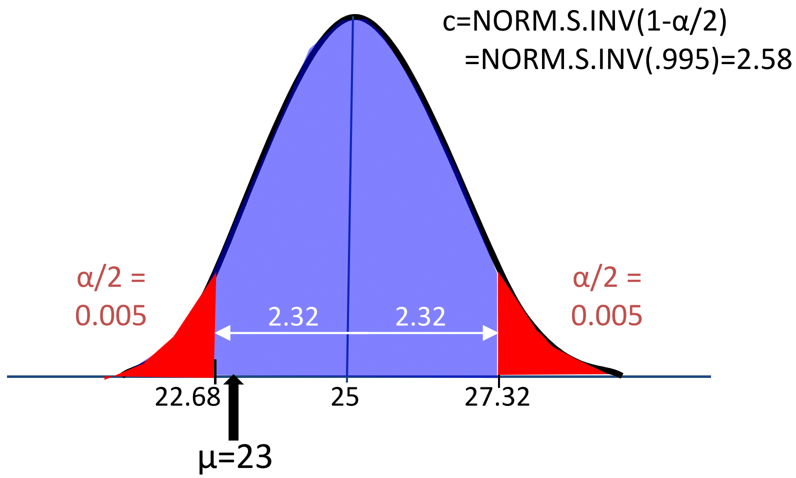
Great! But is this the actual level of significance?

Introduction of p-value

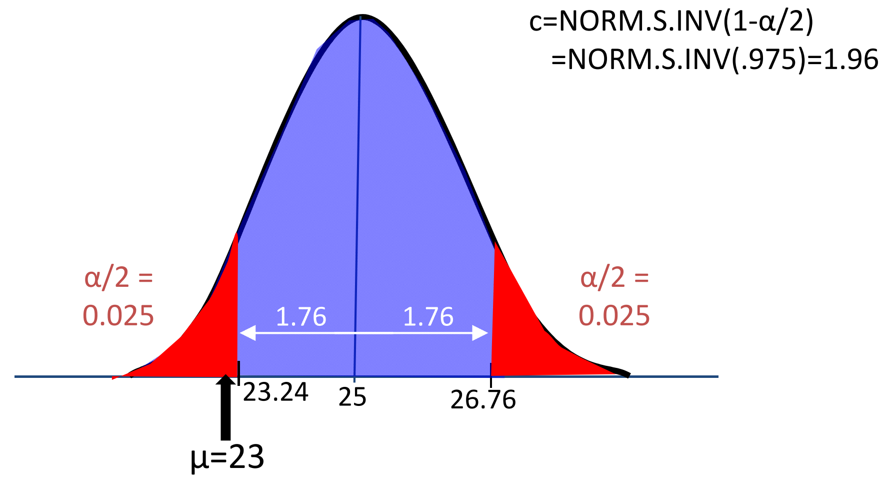
P-Value

p-value The probability, under the assumption of hypothesis, of obtaining a result equal to or more extreme than what was actually observed.

99% Confidence Interval (α=1%)



95% Confidence Interval (α=5%)



At what level of significance, α, does the population mean just fall within the confidence interval?

$$\left[\bar{x} - \frac{cs}{\sqrt{n}},\ \bar{x} + \frac{cs}{\sqrt{n}}\right]$$

In other words, at what confidence value, c, will my population mean be at one of the bounds?

$$\mu = \bar{x} - \frac{cs}{\sqrt{n}}$$

$$c = (\bar{x} - \mu)\left(\frac{\sqrt{n}}{s}\right)$$

c = (25 – 23)/(9/√100)=2.22

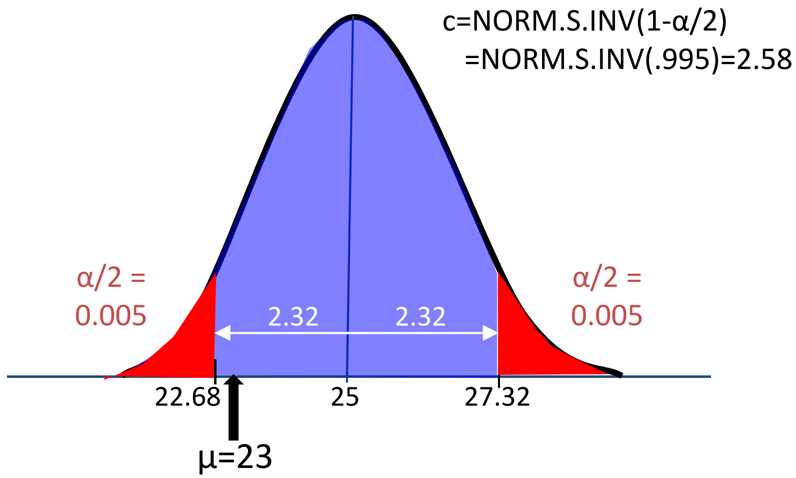
So, What value for α gives me c=2.22?

=NORM.S.DIST(2.22,1) = 0.9868 so α=0.0264 or 2.6%

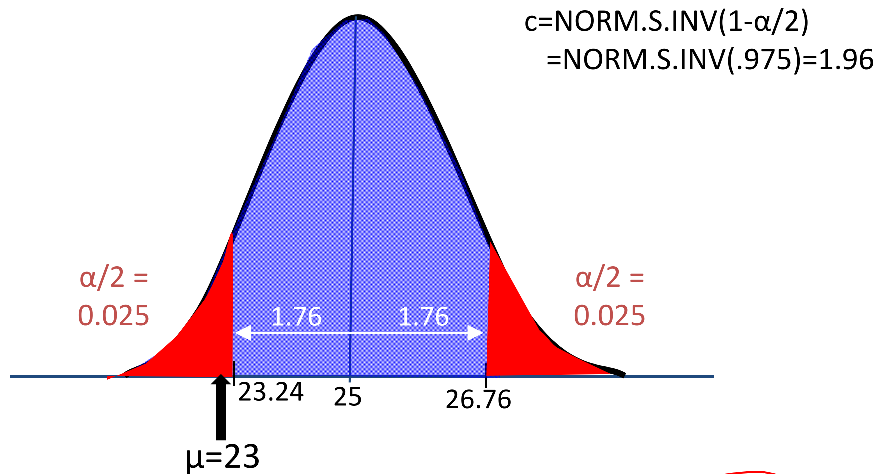
The p-value = 0.026

P-Value

99% Confidence Interval (α=1%)



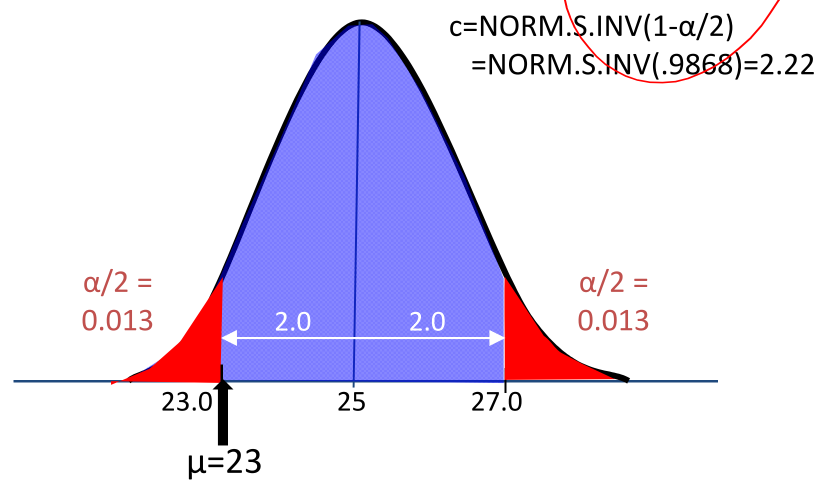
95% Confidence Interval (α=5%)



The p-value tells me the exact level of significance of the point where we would reject H_0. We will use this a great deal in Hypothesis Testing, regression, and elsewhere.

Note that I could have calculated this at the start of my test and known whether I could reject my hypothesis or not!

97.4% Confidence Interval (α=2.6%)



Key Take-Aways

Key Take-Aways (1/2)

- Central Limit Theorem
  - Definition
    - $X_i,..X_n$ are iid with mean=μ and standard deviation = σ
    - The <u>sum</u> of the n random variables is S_n
    - The <u>mean</u> of the n random variables is $\bar{X}$

$$\boxed{S_n = \sum_{i=1}^{N} X_i} \quad \boxed{\bar{X} = \frac{S_n}{N} = \frac{\sum_{i=1}^{N} X_i}{N}}$$

  - Then, if **n** is "large"(say > 30)
    - S_n is Normally distributed with mean = **nμ** and standard deviation **$\sigma\sqrt{n}$**
    - $\bar{X}$ is Normally distributed with mean = μ and standard deviation **$\sigma/\sqrt{n}$**
- Confidence & Prediction Intervals
  - Confidence Interval
    - If we repeat this sampling an infinite number of times, the true mean of this population should appear in 90% of those confidence intervals.
    - Size of sample determines the distribution to use for finding c.

$$\boxed{\left[\bar{x} - \frac{cs}{\sqrt{n}}, \; \bar{x} + \frac{cs}{\sqrt{n}}\right]}$$

  - Prediction Interval
    - If we randomly choose other data points from this population, we would expect 90% of them to fall within this interval.

$$\boxed{\left[\bar{x} - cs, \; \bar{x} + cs\right]}$$

Key Take-Aways (2/2)

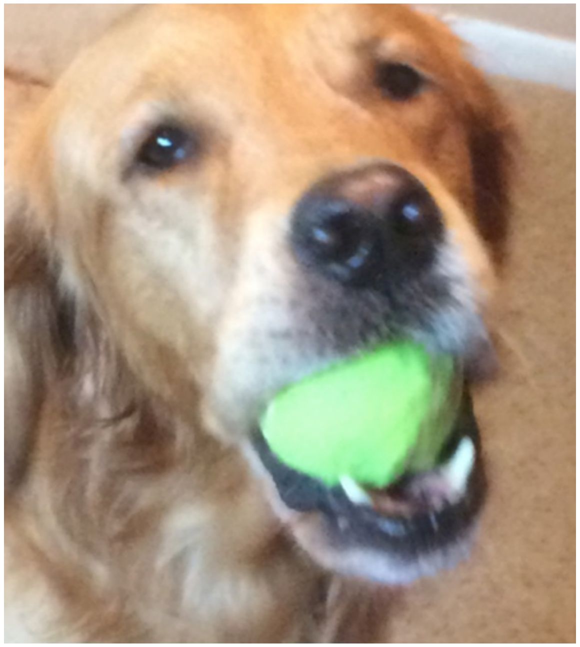
$$\left[\bar{x} - \frac{cs}{\sqrt{n}},\ \bar{x} + \frac{cs}{\sqrt{n}}\right]$$

- Sampling and Sample Size
  - Tradeoffs between interval, sample size, and confidence.
    - If n is fixed, using a higher confidence value c leads to a wider interval L.
    - If c is fixed, increasing sample size n leads to smaller interval L.
    - If both n & c are fixed, reducing variability s leads to smaller interval L.
  - Sample Size
    - We can set n to meet a specific tolerance level L and c.

$$n = \frac{c^2 s^2}{L^2}$$

- Hypothesis Testing
  - Method for making a choice between two mutually exclusive and collectively exhaustive alternatives/hypotheses:
    - Null Hypothesis (H0) - Cannot be rejected unless data argues overwhelmingly otherwise
    - Alternative Hypothesis (H1) The other possible outcomes – mutually exclusive of H0
  - We test, at a specified significance level, to see if we can:
    - Reject the Null Hypothesis, or Fail to reject the Null Hypothesis
    - P-value –the exact level of significance of the point where we would reject H0.

Questions, Comments, Suggestions?
Use the Discussion Forum!



"Wilson testing the hypothesis that someone will throw him the ball."