The Data Driven Manager

5

Making Decisions with Probability Distributions

The Binomial Distribution

Learning Objectives

- Describe the Binomial probability distribution
- Calculate probabilities using the Binomial distribution

The Binomial Distribution

- The Binomial distribution relates to a discrete random variable (nominal data).

- The basis of this distribution is the Bernoulli process.

The Bernoulli Process

- Each trial or experiment has only **two** possible outcomes
- The probability of any and all outcomes remains fixed over time (**constant** probability)
- The trials or experiments are statistically **independent**

The Binomial Formula

$$P(r\ in\ n\ trials) = \left[\frac{n!}{r!(n-r)!}\right][p^r]\left[q^{n-r}\right]$$

where

p = probability of occurrence

q = 1-p = probability of failure

r = number of occurrences desired

n = number of trials

Binomial Example

- A vendor frequently ships 2 bad parts out of 10.
- Suppose the vendor ships our company 50 parts. If we tell them that at least 9 parts out of 10 must be good, and nothing in their manufacturing process has changed, what is the probability that we will receive what we asked for?

Binomial Example

p = 0.80, q = 0.20, r = 45, n = 50

$$P(45\; in\; 50) = \left[\frac{50!}{45!(50-45)!)}\right] [0.8^{45}]\ [0.2^{5}] = 0.02953$$

Binomial Distribution

In RStudio and ROIStat

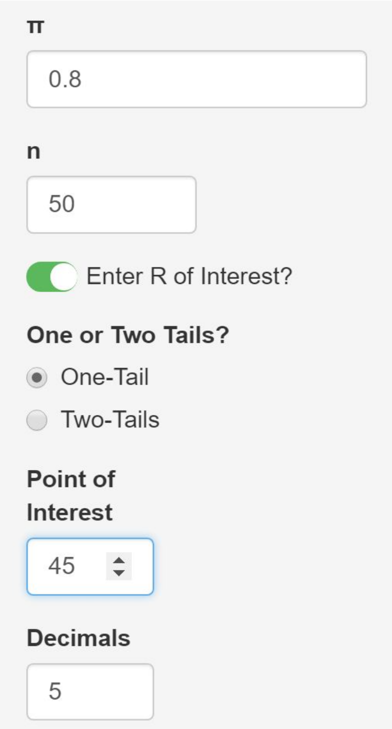
Binomial Example in RStudio

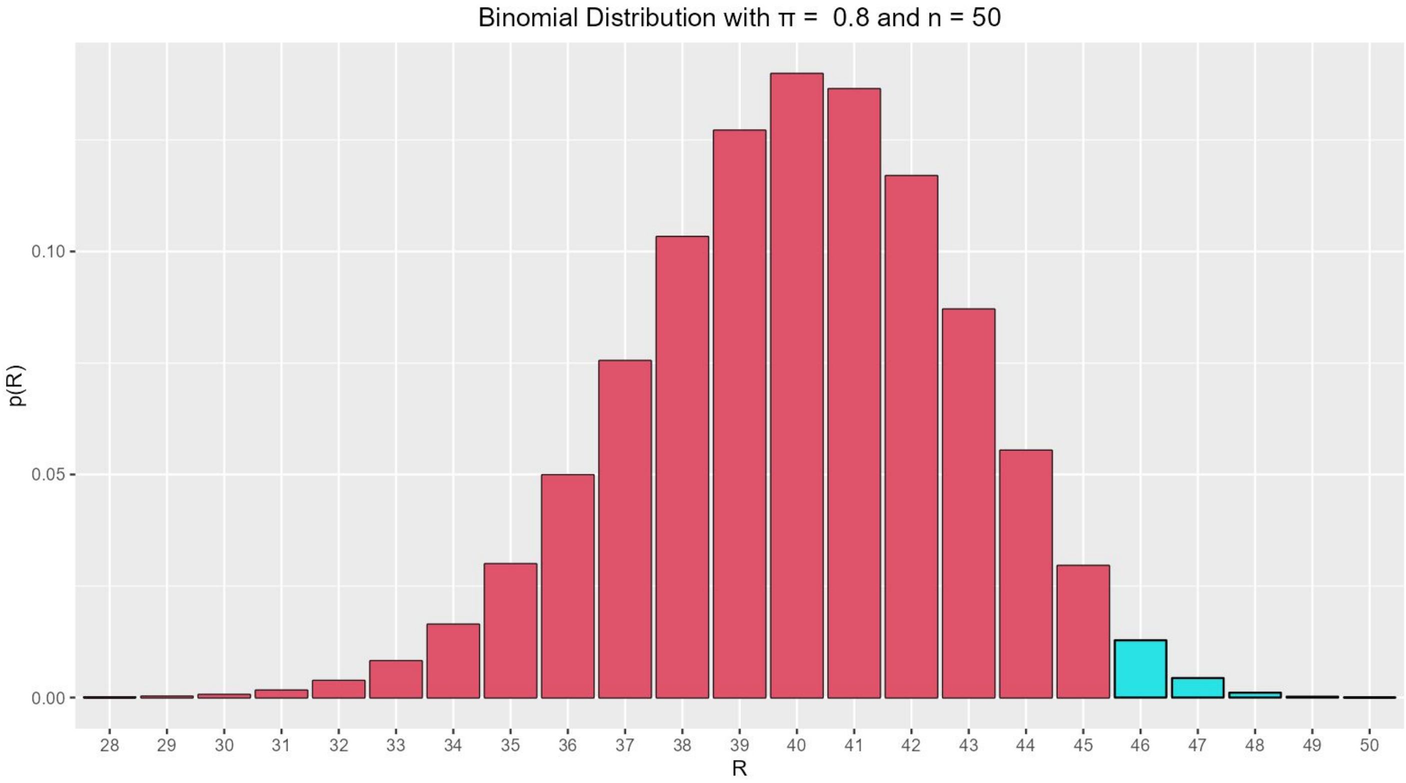
- p = 0.80, q = 0.20, r = 45, n = 50

```
dbinom(x = 45, size = 50, prob = 0.8)
ro(table.dist.binomial(n = 50, p = 0.80),5)
```

Binomial Example in ROIStat

- Open ROI Stat
- Go to Distributions > Binomial
- Enter in the value for p (**π**)
- Enter in the sample size (n)
- Select the Point (R) of Interest





p(45) = 0.02953

p(45 and below) = 0.9815

p(45 and above) = 0.04803

Binomial Example

- What if we wanted to know the probability of getting at least 9 out of 10 good parts in the shipment of 50? $P \geq 45$?
- We would sum the following:

$$P(45) + P(46) + P(47) + P(48) + P(49) + P(50)$$

Binomial Example

- p = 0.80, q = 0.20, r = 45, n = 50

```
# pbinom gives P[X>x] for upper tail
probabilities
pbinom(q = 44, size = 50, prob = 0.80
       , lower.tail = F)
ro(table.dist.binomial(n = 50, p = 0.80),5)
```

The Poisson Distribution

Learning Objectives

- Describe the Poisson probability distribution
- Calculate probabilities using the Poisson distribution

The Poisson Distribution

- This probability distribution is for discrete random variables which can take integer (whole) values (ordinal data)
- Examples:
  - The number of parts produced during a 10 minute period
  - The number of breakdowns per shift
  - The number of failures per 100 cycles

The Poisson Formula

$$P(X) = \frac{\lambda^X}{X!} e^{-\lambda}$$

where

P(X) = probability exactly X occurrences

λ = Mean number of occurrences per time interval (or unit)

e = 2.71828

Poisson Example

λ = 25 parts produced per hour

X = 10 parts produced in one hour

- What is the probability of exactly 10?

$$P(10) = \frac{25^{10}}{10!}e^{-25} = 0.000365$$

Poisson Distribution

In RStudio and ROIStat

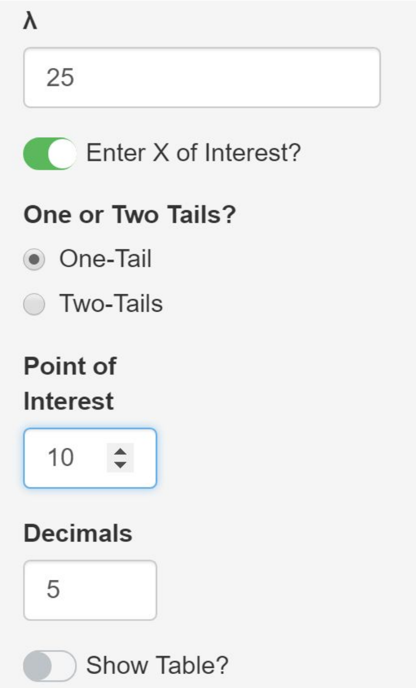
Poisson Example in RStudio

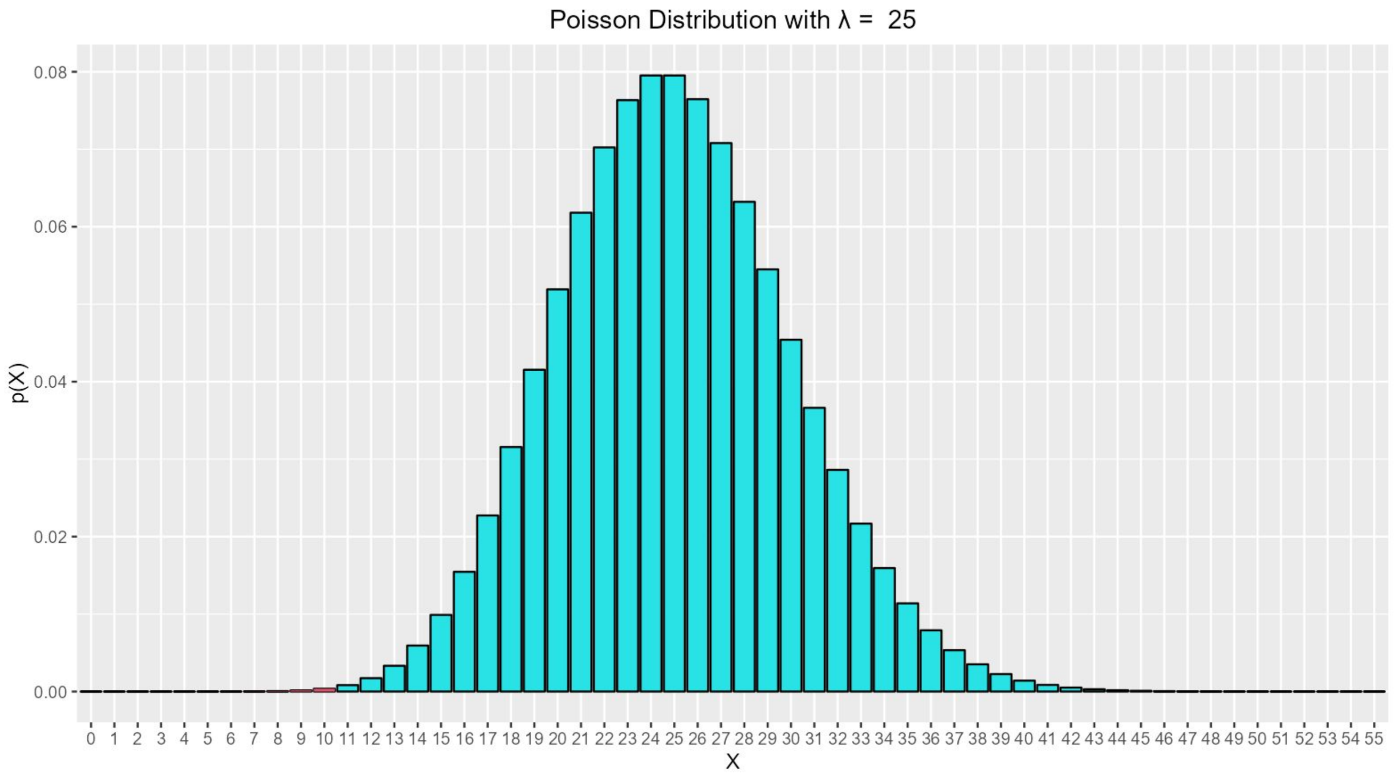
- λ = 25 parts produced per hour
- X = 10 parts produced in one hour
- What is the probability of exactly 10?

```
dpois(x = 10, lambda = 25)
ro(table.dist.poisson(lambda = 25),5)
```

Poisson Example in ROIStat

- Open ROI Stat
- Go to Distributions > Poisson
- Enter in the value for the count (λ)
- Select the Point (R) of Interest





p(10) = 0.00036

p(10 and below) = 0.00059

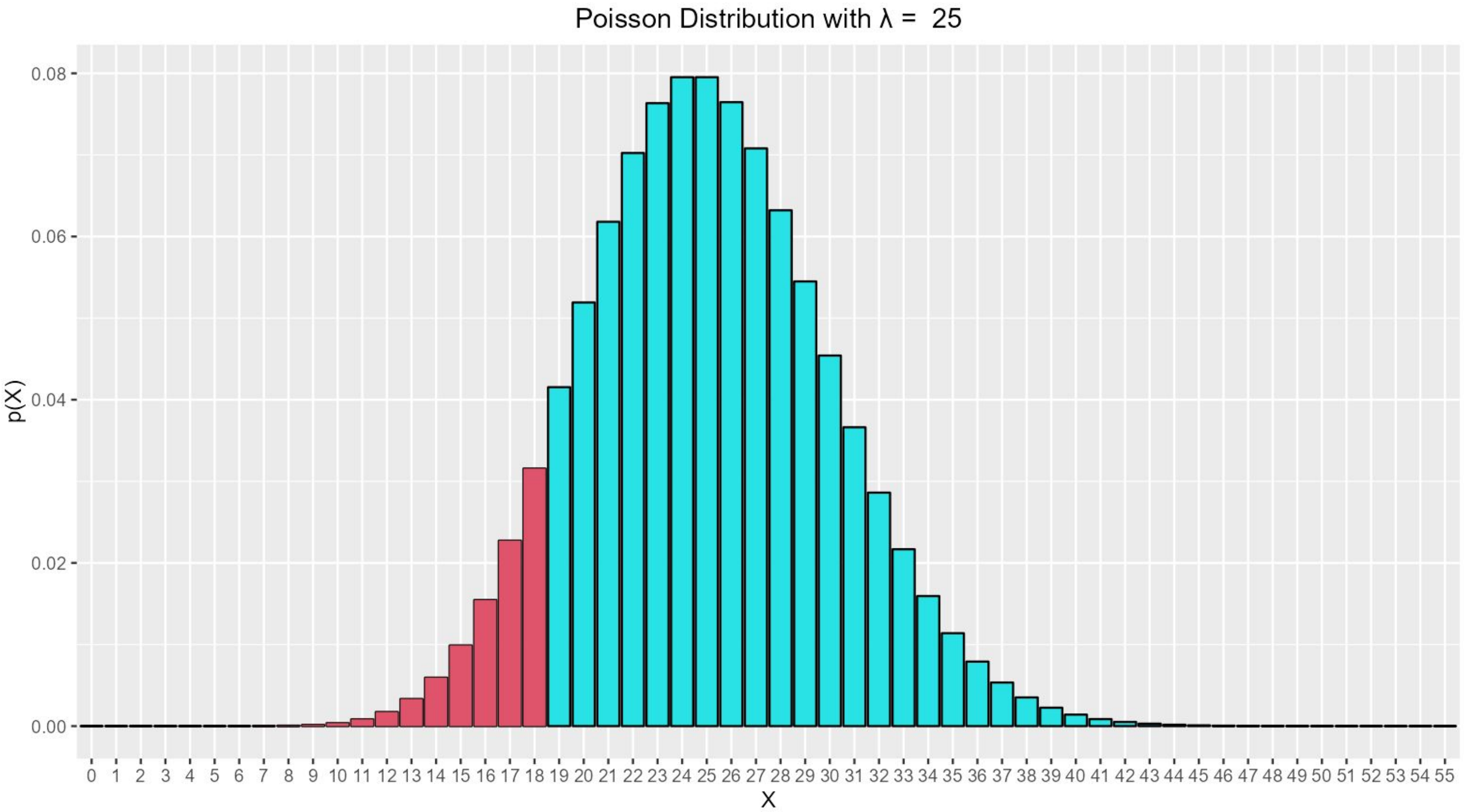
p(10 and above) = 0.99978

Poisson Example

- What is the probability of producing 18 or fewer?
  - λ = 25 parts produced per hour
  - $X \leq 18$ parts produced in one hour

Poisson Example

```
ppois(q = 18, lambda = 25, lower.tail = T)
ro(table.dist.poisson(lambda = 25)[7:51,],5)
```



p(18) = 0.03157

p(18 and below) = 0.09204

p(18 and above) = 0.93952

Testing for a
Poisson Distribution

- It should be noted that all ratio discrete, count data do not necessarily conform to a Poisson Distribution!

- We must ask, therefore, when presented with such sample data set: "Is it reasonable to infer that the data were drawn from a population that may be approximated by a Poisson Distribution?

Testing for a
Poisson Distribution

- Testing in RStudio
  - poisson.dist.test(x = Discrete$DEFECTS)

Testing for a
Poisson Distribution

Although we have not yet discussed it in full, if the p-value is less than 0.05 we reject the hypothesis associated with the test (that is, the data are likely from a Poisson distribution).

Remember this mantra: If p is low, Reject H_0

Testing for a

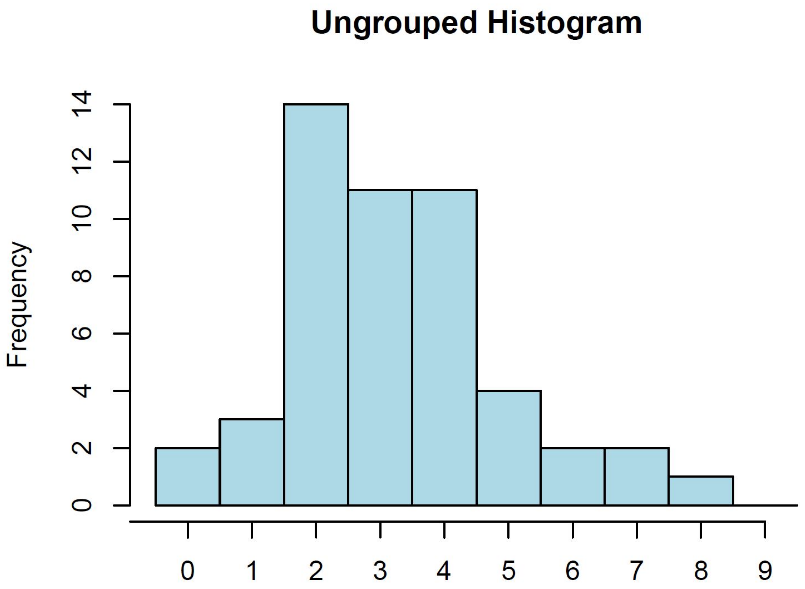
Poisson Distribution

```
> poisson.dist.test(x = Discrete$DEFECTS)

        Poisson Distribution Fit Test Using Variance and Mean

data:  input data
chi.square = 44.173, degrees of freedom = 49, p-value = 0.6624
alternative hypothesis: true chi.square is not equal to 49
sample estimates:
     chi.square sample variance     sample mean
      44.172840        2.920816        3.240000
```

Testing for a

Poisson Distribution

```
> hist.ungrouped(Discrete$DEFECTS)
```



Testing for a

Poisson Distribution

In ROIStat

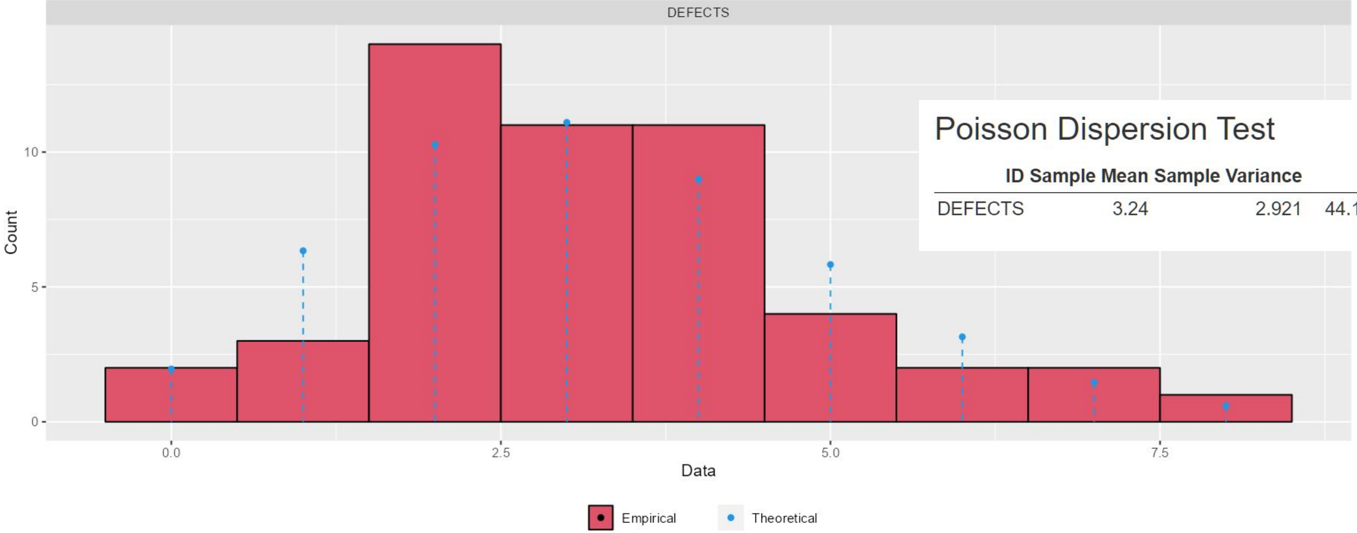
- Open ROI Stat
- Go to Distributions > Testing
- Select the data
- Reject if the p value is < 0.05

Testing for a

Poisson Distribution

In ROIStat

Empirical and Theoretical Histograms



Poisson Dispersion Test

ID	Sample Mean	Sample Variance	χ^2	df	p
DEFECTS	3.24	2.921	44.173	49	0.662

Graphs are only mildly informative. Make your decision using the appropriate test.

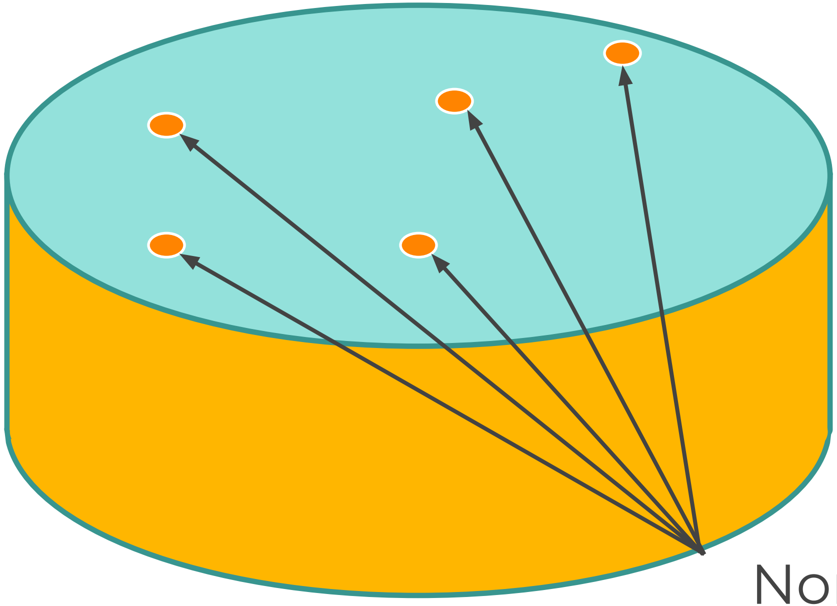
Poisson Distribution

Used for monitoring the number of occurrences of a specified event in a specified inspection unit

Inspection units can be length, area, number of parts, volume, or time.

$$\frac{c}{Unit\ Size}$$

Example - Nonconformities



Nonconformities
c = 5

Thinking **Challenge**

You work in a software development firm as a supervisor. For every **750** lines of code in programs written by a particular software engineer, you know that historically, there will be an average of **6** errors.

Assuming that this engineer has just finished writing an application containing **255** lines of code, what is the probability that this application will be error–free (i.e., have 0 errors)?

Poisson Distribution

Solution: Finding λ

750 lines of code ~ 6 errors = λ

In a 255 line program, we would expect:

$$\lambda = (255 / 750)(6) = 2.04 \text{ errors}$$

or

$$\lambda = 6/750 = 0.008 \text{ errors per line} \times 255 \text{ lines; so}$$

$$0.008 \times 255 = 2.04 \text{ errors}$$

Poisson Distribution Solution: Finding P(0)

Produce the distribution for the relevant Poisson Distribution (λ = 2.04) with the following command, including rounding:

```
round.object(table.dist.poisson(2.04),4)
```

The table is on the next slide.

Poisson Distribution

Solution: Finding P(0)

```
> round.object(table.dist.poisson(2.04),4)
    x p.at.x eq.and.above eq.and.below
0   0 0.1300       1.0000       0.1300
1   1 0.2653       0.8700       0.3953
2   2 0.2706       0.6047       0.6659
3   3 0.1840       0.3341       0.8498
4   4 0.0938       0.1502       0.9437
5   5 0.0383       0.0563       0.9819
6   6 0.0130       0.0181       0.9950
7   7 0.0038       0.0050       0.9988
8   8 0.0010       0.0012       0.9997
9   9 0.0002       0.0003       0.9999
10 10 0.0000       0.0001       1.0000
11 11 0.0000       0.0000       1.0000
```

Discrete Probability Distributions

Practice Activities

Binomial Distribution

Example:

- Assume a supplier has a consistent 10% nonconforming rate. Suppose that the supplier ships 50 parts to your plant in a single lot.
- What is the probability of finding **exactly two** nonconforming parts in the 50 parts?
- What is the probability of finding **two or less** nonconforming parts in the 50 parts?

Binomial Distribution

Example:

You can use lolcat's '`table.dist.binomial()`' function considering:

$\pi = 0.10$

$n = 50$

$r = 2$

Binomial Distribution

```
> ro(table.dist.binomial(n,p)[1:10,],4)
  x p.at.x eq.and.above eq.and.below
0 0 0.0052       1.0000       0.0052
1 1 0.0286       0.9948       0.0338
2 2 0.0779       0.9662       0.1117
3 3 0.1386       0.8883       0.2503
4 4 0.1809       0.7497       0.4312
5 5 0.1849       0.5688       0.6161
6 6 0.1541       0.3839       0.7702
7 7 0.1076       0.2298       0.8779
8 8 0.0643       0.1221       0.9421
9 9 0.0333       0.0579       0.9755
```

The exact probability of x, or r = 2 can be obtained with the following R function:

```
dbinom(x = 2,size = 50, prob = 0.1)
```

The probability of 2 or fewer can be obtained with the following R function:

```
pbinom(q = 2, size = 50, prob = 0.1)
```

Binomial Distribution

- Assume that a product has a documented failure rate of **0.20** after 150 hours of use. If we were to place **30** randomly selected parts from this process in the field:

- π = ______
- n = ______

Binomial Distribution

- Assume that a product has a documented failure rate of 0.20 after 150 hours of use. If we were to place 30 randomly selected parts from this process in the field:
  - What is the probability that **5 or fewer** will have failed?

Binomial Distribution

- Assume that a product has a documented failure rate of 0.20 after 150 hours of use. If we were to place 30 randomly selected parts from this process in the field:
    - What is the probability that **exactly 5** will have failed after 150 hours?

Binomial Distribution

- Assume that a product has a documented failure rate of 0.20 after 150 hours of use. If we were to place 30 randomly selected parts from this process in the field:
  - What is the probability that **more than 10** will have failed?

Poisson Distribution

Example:

- The number of OSHA-recordable safety accidents in a manufacturing plant has been running 4.2 accidents per 200,000 hours worked. What is the probability of having exactly two accidents in a 200,000-hour work period?

- Given, λ = 4.2, X=2
- P(2) = ___________________

Poisson Distribution

Example:

- You can use lolcat's '`table.dist.poisson()`' function to get the results (next slide) or directly with the R `dpois()` function, both demonstrated on the next slide:
- λ = 4.2
- X = 2

Poisson Distribution

```
ro(table.dist.poisson(lambda)[1:5,],4)
  x p.at.x        eq.and.above        eq.and.below
0 0 0.0150              1.0000              0.0150
1 1 0.0630              0.9850              0.0780
2 2 0.1323              0.9220              0.2102
3 3 0.1852              0.7898              0.3954
4 4 0.1944              0.6046              0.5898
```

Poisson Distribution

Example:

- An expeditor has been monitoring the daily production rate of blanked saw chain cutters. On average, the number of buckets per day that have been produced is 65 (**λ**) and the output is representative of a Poisson function
- What is the probability of producing 50 buckets or more in a day?

The Normal Distribution

Learning Objectives

- Describe the Normal probability distribution
- Calculate probabilities using the Standard Normal distribution

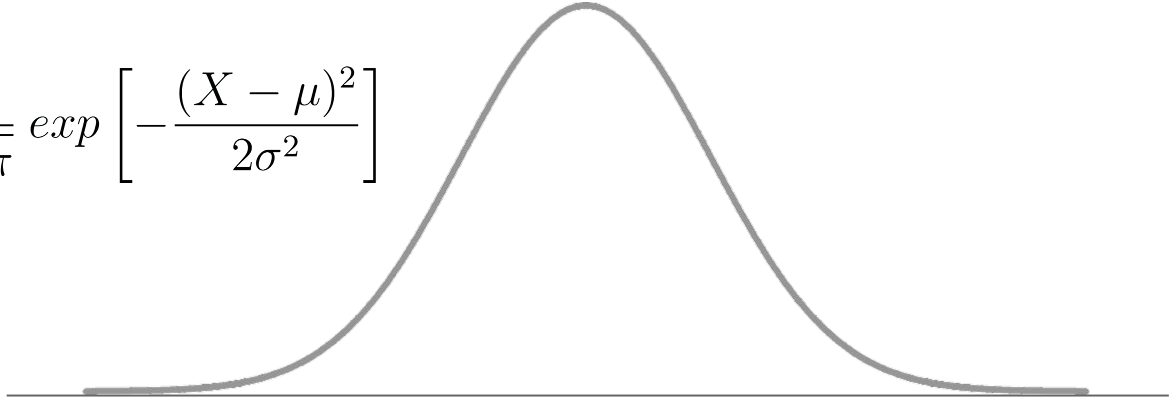
The Normal Distribution

- A theoretical probability distribution for a continuous random variable
- One of the most important distributions because of its wide range of practical applications

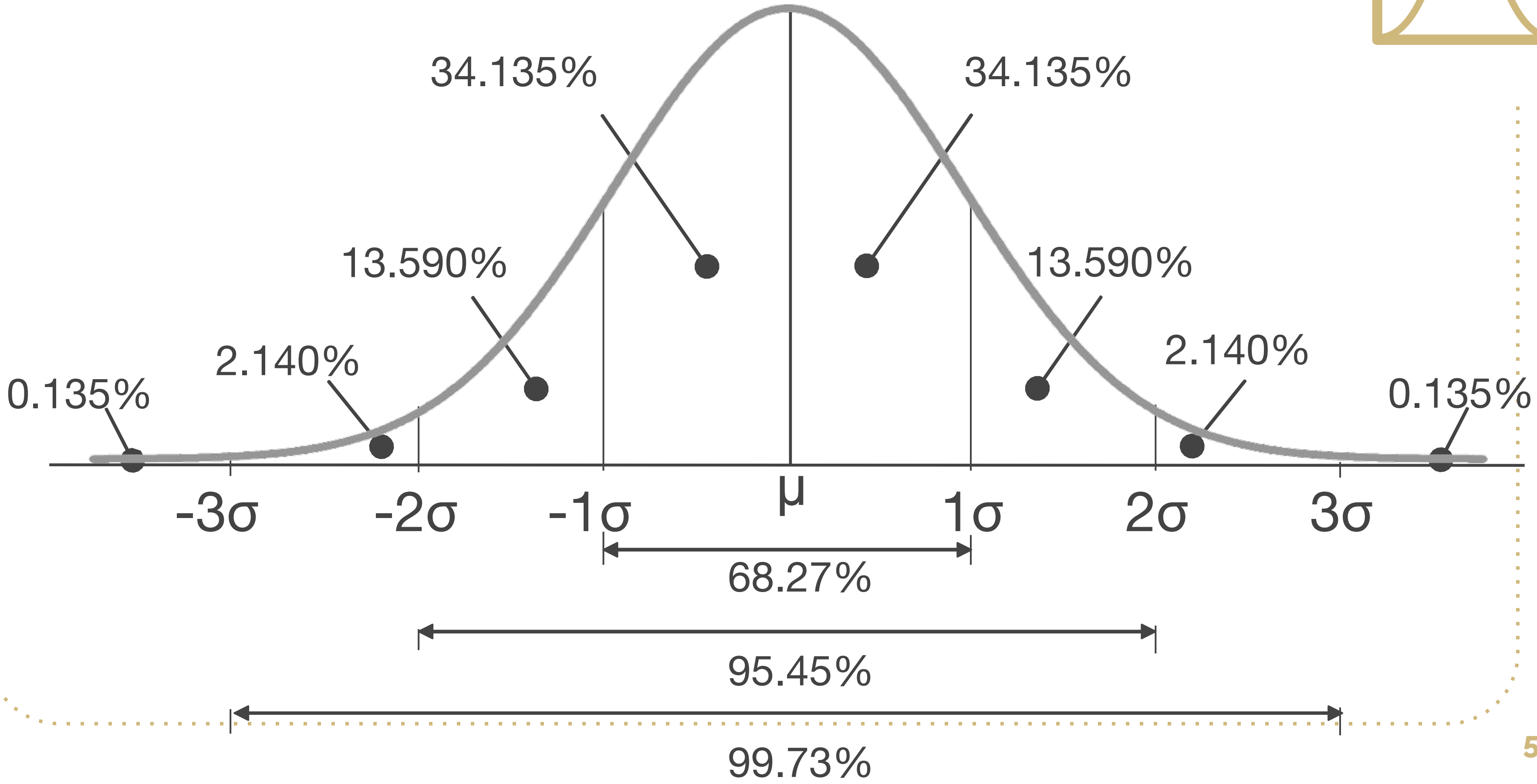
The Normal Distribution

- Mean = Median = Mode
- Symmetrical around μ
- Tails extend to ∞ but never touch the horizontal axis
- $\gamma_3 = 0.00$
- $\gamma_4 = 0.00$
- Areas under the curve are predictable

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right]$$



Areas Under the Normal Curve



Area Calculations

- The area corresponding to any score value may be found using a z-score, where

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

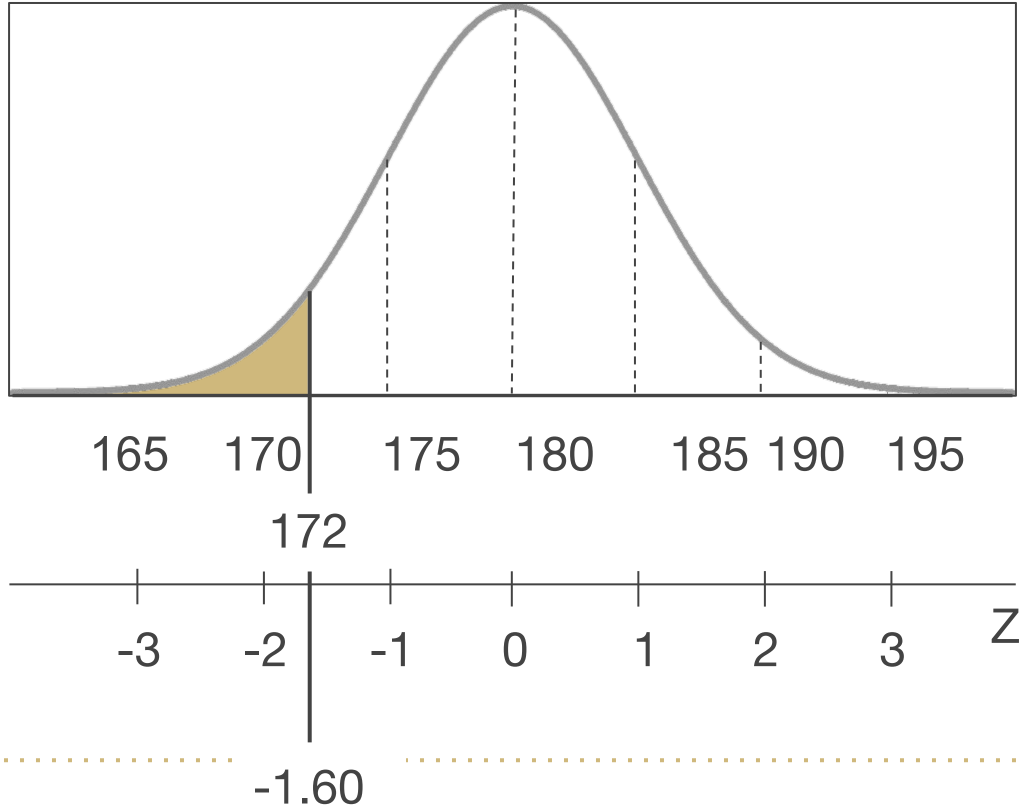
- Z is the number of standard deviation units from X to μ

Normal Distribution Example 1

- To date, tooling used on a particular drilling process has lasted an **average of 180 hours** before requiring replacement, with a **standard deviation of 5 hours**.

- What is the probability that a tool selected at random from the tool crib will last **less than 172 hours** before replacement is required?

$$Z = \frac{X - \mu}{\sigma} = \frac{172 - 180}{5} = -1.60$$

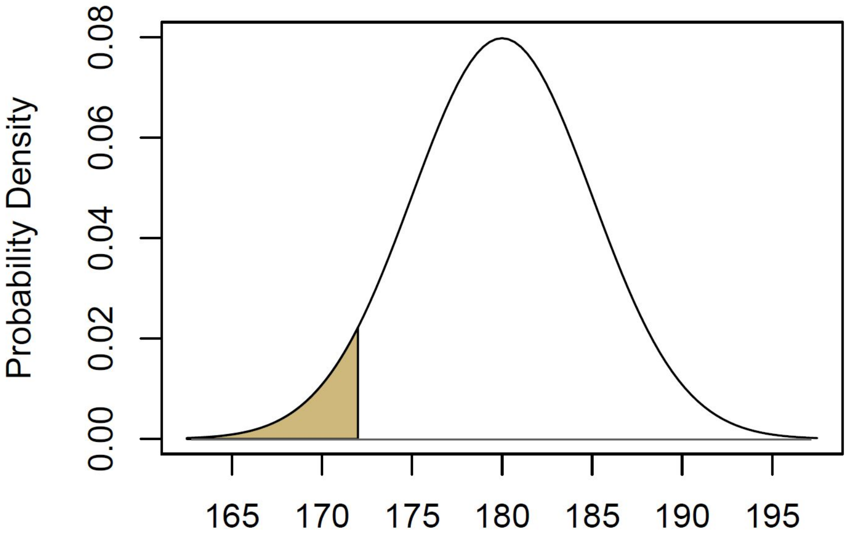


Normal Distribution

In RStudio and ROIStat

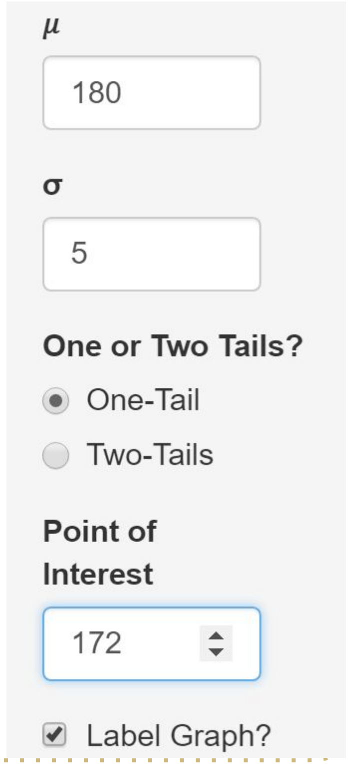
Normal Distribution in RStudio

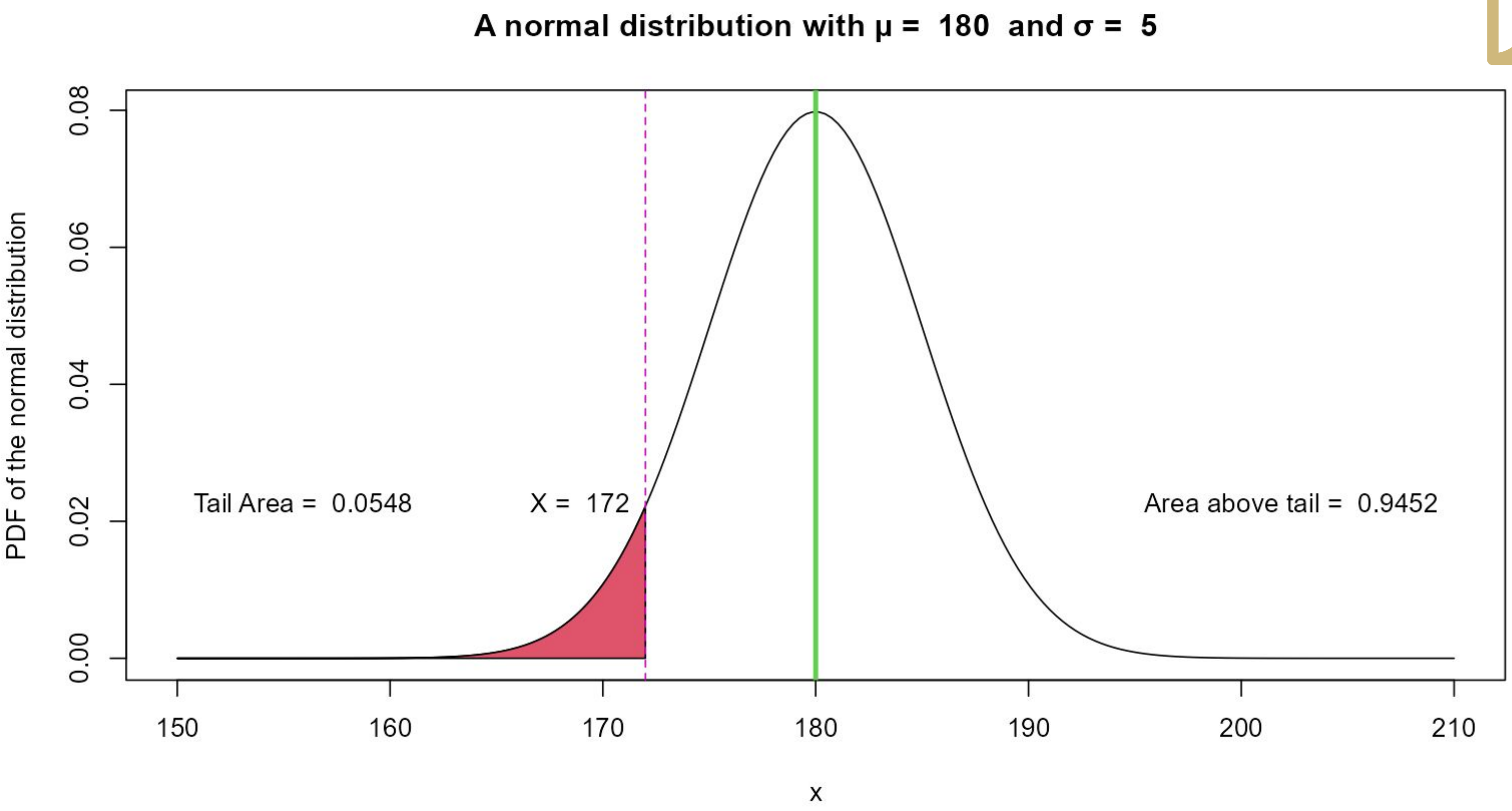
```
> pnorm(q, mean, sd
        , lower.tail)
```



Normal Distribution in ROIStat

- Open ROI Stat
- Go to Distributions > Normal
- Enter in the value for the average (**μ**)
- Enter in the value for the std. dev. (**σ**)
- Select the Point of Interest




A normal distribution with μ = 180 and σ = 5
PDF of the normal distribution
0.00
0.02
0.04
0.06
0.08
Tail Area = 0.0548
X = 172
Area above tail = 0.9452
150
160
170
180
190
200
210
x

Normal Distribution Example 2

- A stamping operation has been running consistently, punching two holes in sheet metal.
- The center-to-center distance between the two holes has been an average (μ) of 5.20mm, with a standard deviation (σ) of 0.05mm.

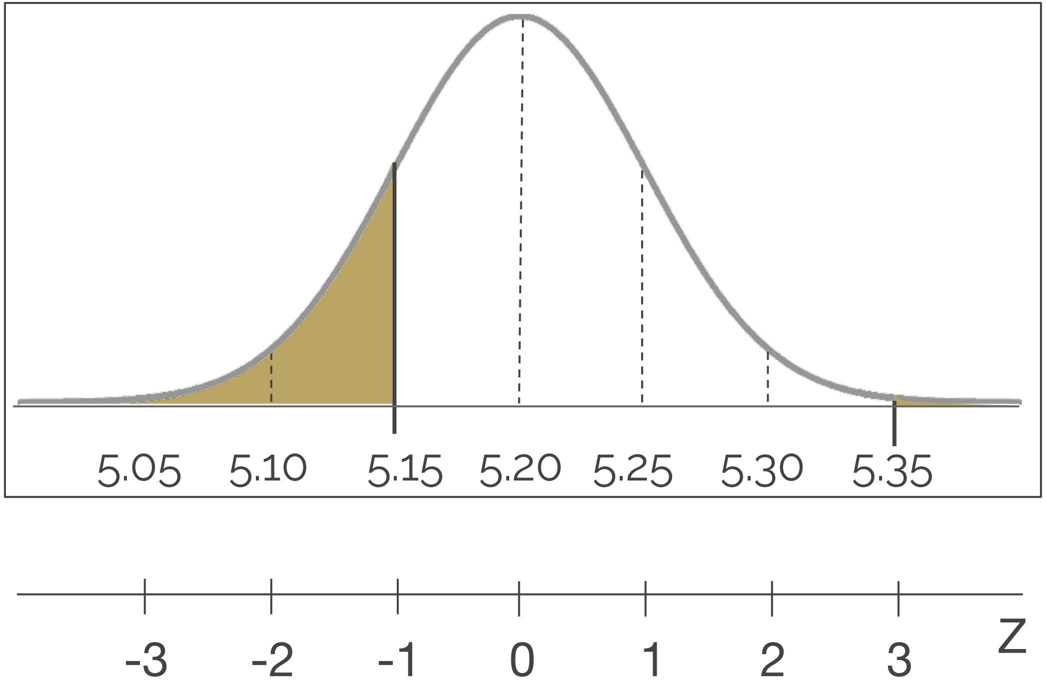
Normal Distribution Example 2

- The process produces center-to-center distances that can be modeled with a normal distribution.
- The specifications for these parts require a maximum, or upper (USL), limit of 5.35mm and a minimum, or lower (LSL), limit of 5.15mm.
- What percentage of the manufactured parts are likely to fall outside of the specifications?

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

$$z = \frac{5.15 - 5.20}{.05}$$

$$z = \text{-}1.00$$



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

$$z = \frac{5.35 - 5.20}{.05}$$

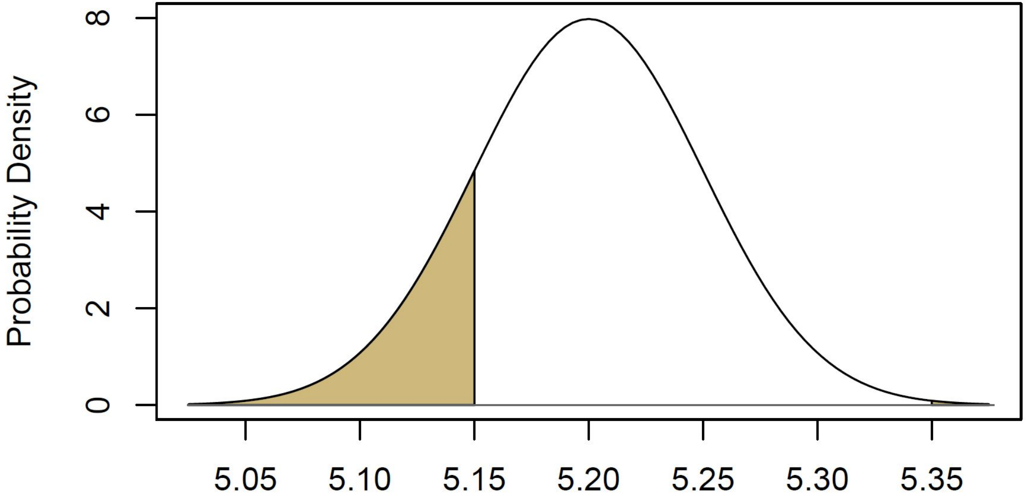
$$z = 3.00$$

Normal Distribution Example 2

- The process produces center-to-center distances that can be modeled with a normal distribution.
- The specifications for these parts require a maximum, or upper (USL), limit of 5.35mm and a minimum, or lower (LSL), limit of 5.15mm.
- What percentage of the manufactured parts are likely to fall outside of the specifications?

Normal Distribution in RStudio

```
> pnorm(q, mean, sd
        , lower.tail)
```



Normal Distribution in ROIStat

- Open ROI Stat
- Go to Distributions > Normal
- Enter in the value for the average (μ)
- Enter in the value for std. dev. (σ)
- Select the Point of Interest

Testing for Normality

- When $n < 25$, use the Anderson-Darling / Shapiro-Wilk tests for normality

- When $n \geq 25$, use Skewness Test, and Kurtosis Test (Moment Tests)

Testing for Normality

- Probabilities ≥ 0.05 indicate that the data are normal
- Probabilities < 0.05 indicate that the data are NOT normal

Testing for Normality in RStudio

In R / Rstudio:

- anderson.darling.normality.test( )
- shapiro.wilk.normality.test( ) or
- summary.continuous( )

Testing for Normality in ROIStat

- Open ROI Stat
- Go to EDA > Normality Tests

OR

- Go to Distributions > Testing

Comparing Actual Out of Spec to Predicted Out of Spec

When calculating the percent out of specification (or above / below a score value), why don't we just count the number of values in the sample?

Comparing Actual Out of Spec to Predicted Out of Spec

Which is more correct? The percentage in the ***sample*** you took, **or** what is ***predicted in the population*** based on the normal distribution (given that we tested for normality and can show that it is probable that the sample was drawn from a normal distribution)?

We want to make an ***inference*** from the *sample* to the **population**!

Comparing Actual Out of Spec to Predicted Out of Spec

Sample - actual out of specification in the sample

```
> sum(data < x)/n or sum(data > x)/n
```

Population - estimated out of specification in the population

```
> pnorm(x, mu, sigma)
```

Comparing Actual Out of Spec to Predicted Out of Spec

Example: Using the FlowRate.txt data file…

What percentage of values ***in the sample*** are < 15?

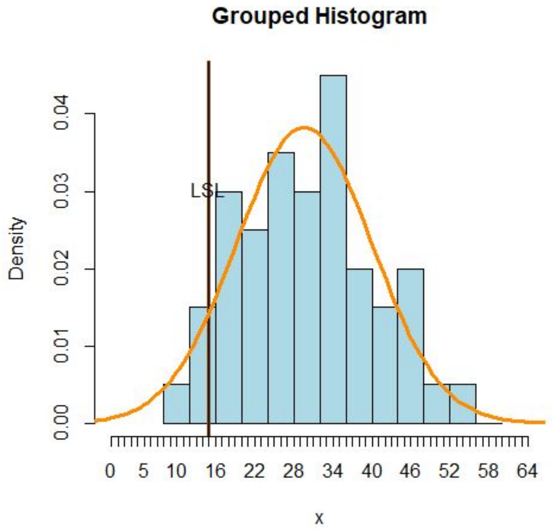
```
> sum(FlowRate$Flow < 15)/50
= 4.00%
```

Comparing Actual Out of Spec to Predicted Out of Spec

What percentage of values ***in the population*** are predicted to be < 15?

```
pnorm(q = 15, mean =
mean(FlowRate$Flow),
sd = sd(FlowRate$Flow),
lower.tail = T)
```

= 8.08%



The Exponential Distribution

Learning Objectives

- Describe the Exponential probability distribution
- Calculate probabilities using the Exponential distribution

The Exponential Distribution

- The exponential distribution occurs in a number of situations in the industrial environment.
- Time to failure often follows an exponential distribution.

The Exponential Distribution

- Measurement from a physical process that has a restraint, such as the location of a hole from a reference edge, where the reference edge is pressed against a fixture, may follow an exponential distribution.

- Roundness of shaft, measured by total indicator reading, may also follow this type of distribution.

The Exponential Distribution

- The exponential distribution is a continuous random variable probability distribution with the form:

$$y = \frac{1}{\mu - X_{min}} e^{\left[-\frac{X - X_{min}}{\mu - X_{min}}\right]}$$

The Exponential Distribution

- When $X_{min} = 0$, the equation reduces to:

$$y = \frac{1}{\mu} e^{\left[-\frac{x}{\mu}\right]}$$

The Exponential Distribution

- The normal distribution contains an area of 50% above and 50% below μ.

- With the exponential distribution, 36.8% of the area under the curve is above the average (μ) and 63.2% is below.

Applications / Observations

- Predictions based on an exponentially distributed process often only require the μ (and sometimes X_{min}) of the process.
- For prediction purposes, finding the area under the curve beyond the time period of concern is generally the point of interest.
- These prediction often relate to reliability issues or time between failure analyses.

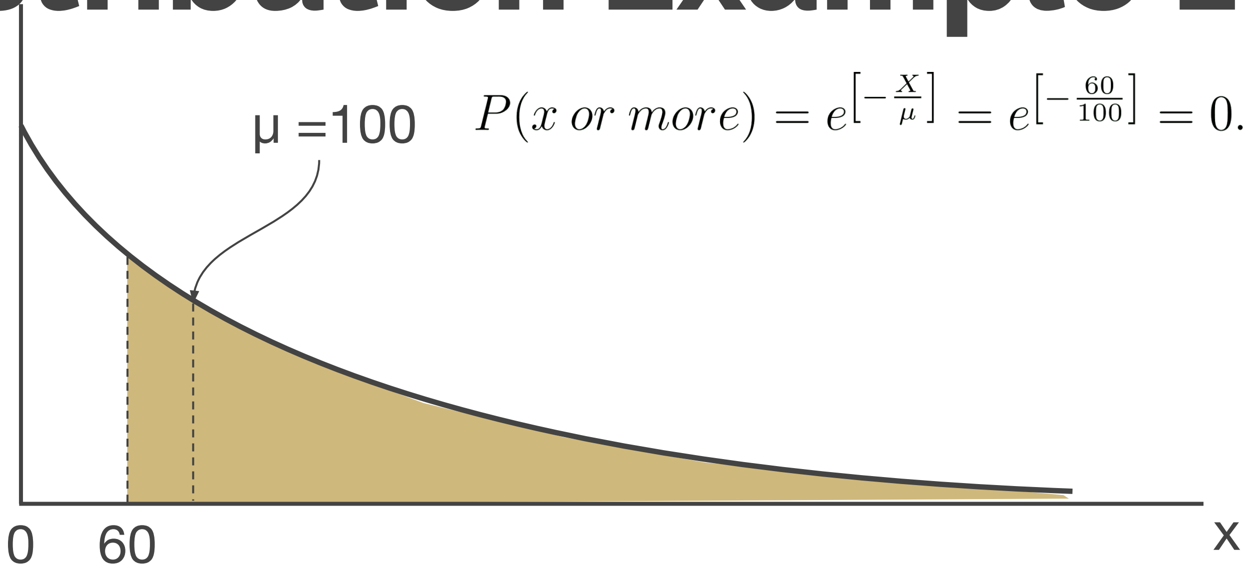
Exponential Distribution Example 1

- An in-plant study has shown that an engine control module laboratory tester is capable of operating on an **average of 100 hours** between breakdowns (MTBF).

- What is the probability that the tester will run for **at least 60 successive hours** without a breakdown (assuming that the time to failure pattern is distributed exponentially)?

Exponential Distribution Example 1

$$P(x\ or\ more) = e^{\left[-\frac{X}{\mu}\right]} = e^{\left[-\frac{60}{100}\right]} = 0.5488$$



Exponential Distribution in RStudio

```
> pexp(q, rate, lower.tail)
```

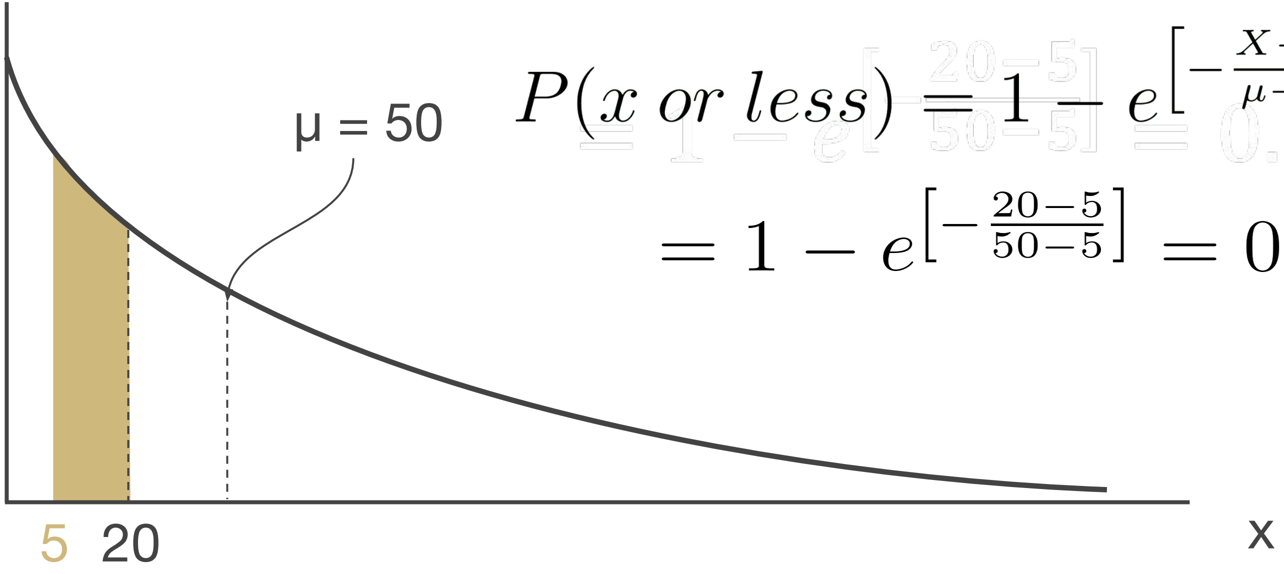
Exponential Distribution in ROIStat

- Open ROI Stat
- Go to Distributions > Exponential
- Enter in the value for the the average (μ)
- Enter in the value for the minimum value (Xmin)
- Select the Point of Interest

Exponential Distribution Example 2

- The distribution of time for a particular grinding machine is characterized by the exponential distribution.
- The mean time between breakdowns has been established at 50 minutes.
- The origin parameter (X_{min}) is estimated to be 5 minutes.
- What is the probability of this machine running 20 minutes or less before a breakdown?

Exponential Distribution Example 2



$$P(x \; or \; less) = 1 - e^{\left[-\frac{X - X_{min}}{\mu - X_{min}}\right]}$$

$$= 1 - e^{\left[-\frac{20-5}{50-5}\right]} = 0.2835$$

Exponential Distribution in RStudio

```
> pexp(q, rate, lower.tail)

> pexp.low(q, low, mean, lower.tail)
```

Exponential Distribution in ROIStat

- Open ROI Stat
- Go to Distributions > Exponential
- Enter in the value for the the average (μ)
- Enter in the value for the minimum value (Xmin)
- Select the Point of Interest

Testing for Exponentiality

- Always test for normality first!
- When $n \leq 100$, use the Shapiro-Wilk test
- When $n > 100$, use the Epps and Pulley test

Testing for Exponentiality

- Probabilities ≥ 0.05 indicate that the data are exponential

- Probabilities < 0.05 indicate that the data are NOT exponential

Testing for Exponentiality in RStudio

```
> shapiro.wilk.exponentiality.test( )

> shapetest.exp.epps.pulley.1986( )
```

Testing for Exponentiality in ROIStat

- Open ROI Stat
- Go to Distributions > Testing > Exponential
- Select the data to test
- If using Shapiro Wilk or MVP, click on the 'Start Simulation' button

Continuous Probability Distributions

Practice Activities

Normal Distribution

Example

Past participants in a training program designed to upgrade the skills of production-line supervisors spent an average of 500 hours on the program, with standard deviation of 100 hours. Assume a normal distribution.

- What is the probability that a participant selected at random will require **more than 500 hours** to complete the program?
- What is the probability that a candidate selected at random will take **between 550 and 650 hours** to complete the program?

Normal Distribution

Example

What is the probability that a candidate selected at random will take between 550 and 650 hours to complete the program?

```
> pnorm(650,500,100) = 0.9331928 # 650 hours or less
> pnorm(550,500,100) = 0.6374625 # 550 hours or less
```

The difference between the two is the answer: 0.2417333

Normal Distribution
Practice Activity

A process has typically run at a **μ** of 163 with a **σ** of 12. The specifications for the part are 169 ± 5.

- What is the probability that a single part selected at random from a standard lot will be out of specification assuming that a normal distribution has been documented?

A = 17.9659%
B = 53.3207%
C = 99.7300%
D = 71.2866%

Exponential Distribution

Example

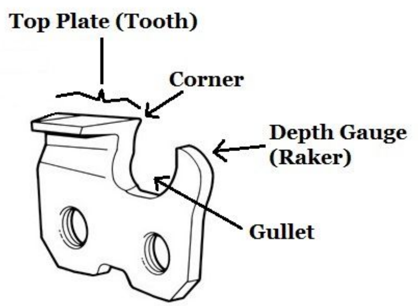
A research study has shown that the time required to return a response to a bid request at an automotive supplier is exponentially distributed with a mean of 72.5 hours; and an origin parameter of 25 hours.

- What percentage of responses are submitted in less than 48 hours ?

Exponential Distribution

Practice Activity

- The distribution of time between breakdowns or stoppages for a particular grinding machine is characterized by the exponential distribution. The grinding machine automatically grinds the cutting edge in the gullet of a saw chain cutter. Statistically, the mean time between breakdowns has been established as 46 minutes. Also, the minimum value is estimated to be five minutes.



A Modern Cutter: Combination of Tooth and Raker

Exponential Distribution

Practice Activity

- What is the probability of this particular machine running 15 minutes or less before a breakdown?

A = 78.3564%
B = 21.6436%
C = 72.1742%
D = 27.8258%

Exponential Distribution
Practice Activity

- What is the probability of it running 60 minutes or more before a failure occurs?

A = 72.8651%
B = 26.1463%
C = 73.8537%
D = 27.1349%