Chapter 3
Uncover Pattern
Blurred by Sampling Instability

Section 3.2
Normal and Related Distributions

David W. Gerbing

The School of Business
Portland State University

- Normal and Related Distributions
  - The Normal Curve
  - Standard Scores
  - Normal Curve Probabilities
  - General Density Curves

3.2a
The Normal Curve

Descriptions of Hypothetical Populations

Build a model of reality for a continuous variable

- Mathematicians have constructed probability distributions of continuous variables from equations that can
  - Meaningfully describe aspects of reality
  - Become the basis for the analyses of inferential statistics
- **Normal curve**: The mathematically defined bell-shaped graphical representation of the family of normal distributions[1]
- The population parameters of a specific normal curve are its
  - Population mean, μ (mu): location of the center
  - Population standard deviation, σ (sigma): dispersion about the center

[1] A normal curve approximates a "smoothed out" binomial distribution. A normal distribution is the limit of a corresponding binomial distribution as the number of trials increases indefinitely.

Normal Curve Attributes: Illustration

Two Normal Curves

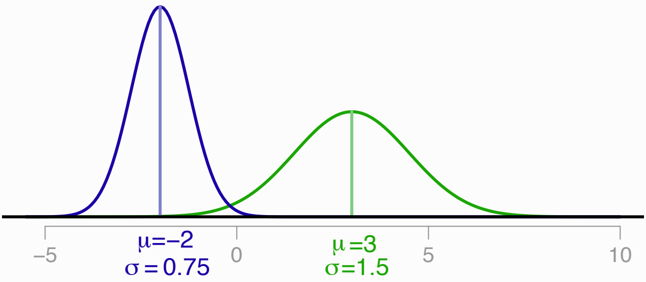


Figure: Two normal curves, with different values of μ and σ

- The green normal curve (on the right) is centered over $\mu = 3$ and the blue normal curve is centered over $\mu = -2$
- The green normal curve, with $\sigma = 1.5$, is wider than the blue normal curve, with $\sigma = 0.75$

Relation of the Normal Curve to Data

The ideal versus the observed

- **Key Concept**: The shape of a histogram of data sampled from a normal population only *approximates* the shape of a normal curve
- The quality of the approximation depends on
  - The bin widths of a histogram can be no smaller than the unit of measurement, which is necessarily larger than the abstraction of a mathematical point of zero width
  - Sampling error: Attributes of a sample, such as its shape, do not perfectly match the corresponding characteristic of the population from which the sample was drawn
- Next illustrate this approximation of data to its underlying perfect form by using the computer to generate simulated samples of data from a known, specified normal distribution

Simulated Samples from a Normal Distribution

"Make data"

- **Monte Carlo Simulation**: A random sample of data generated by the computer according to a specified probability distribution, a population, such as the normal distribution
- Monte Carlo data can be used to
  - illustrate known statistical principles
  - discover how a statistic performs in specific situations
- Why simulate samples of data? Why is simulation important?
- The answer is that every single data analysis is of data sampled from at least one population
- **Key Concept**: Understanding how the sample of data relates to the population from which it is sampled is a central pursuit of the entire enterprise of data analysis
- Here examine a consequence of sample size

Simulated Samples from a Normal Distribution

R can "Make data"

- To simulate n normal data values with R and then store the results in a vector, here named Y, use the `rnorm` function

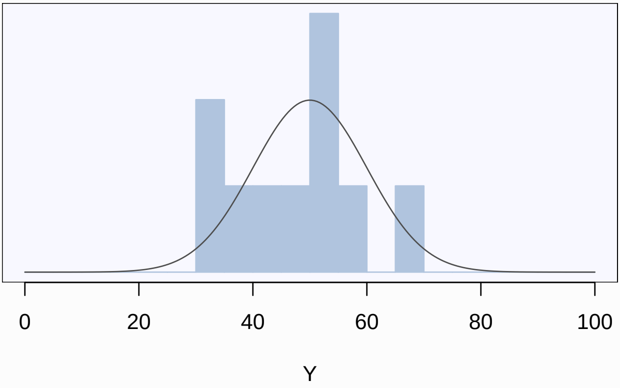
```
> Y <- rnorm(n,mean,sd)
```

- The specified values of `mean` and `sd` set the corresponding population values of a specific normal distribution, μ and σ
- To list the generated values, enter the vector name: `> Y`
- Can analyze, such as with the following call to `Histogram()` with `density=TRUE` that generates the following figures

  $>$ Histogram(Y, density=TRUE)

Histogram #1 of Data from Normal Population, $n = 10$

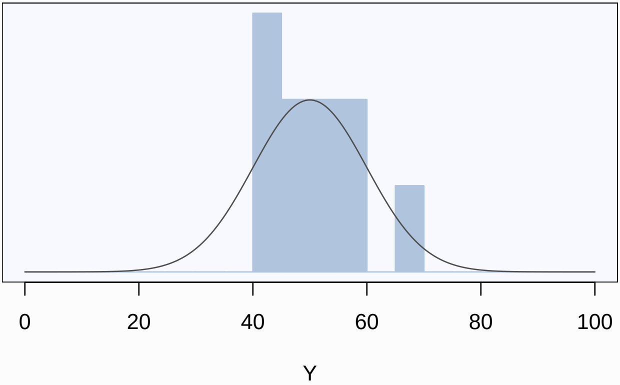
```
> Y <- rnorm(n=10, mean=50, sd=10)
```



- This histogram of 10 random data values from a normal population only roughly approximates a normal distribution

Histogram #2 of Data from Normal Population, $n = 10$

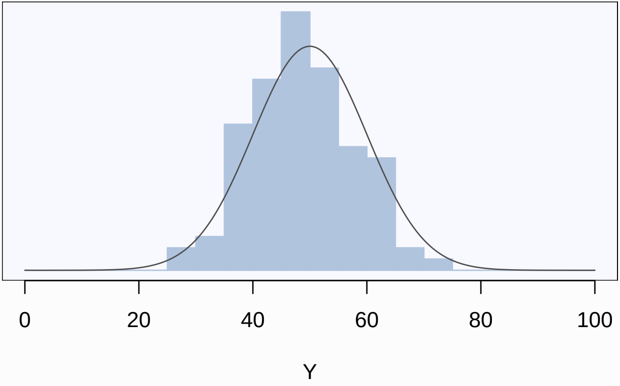
```
> Y <- rnorm(n=10, mean=50, sd=10)
```



- This histogram of 10 different random data values from a normal population only vaguely resembles the previous distribution

Histogram of Data from Normal Population, $n = 100$

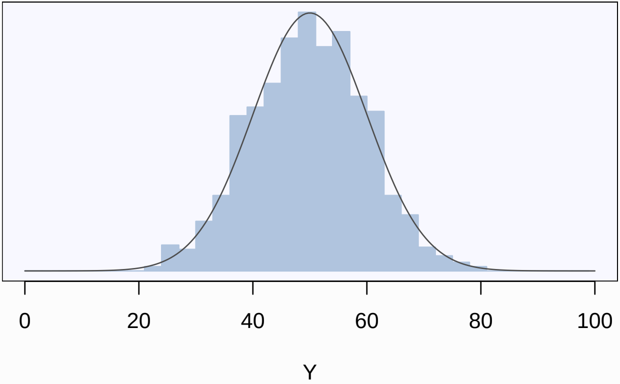
```
> Y <- rnorm(n=100, mean=50, sd=10)
```



- This histogram of 100 random data values from a normal population only roughly approximates a normal distribution

Histogram of Data from Normal Population, $n = 1000$

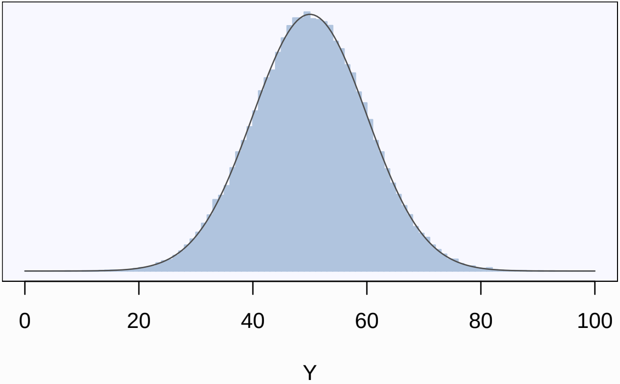
```
> Y <- rnorm(n=1000, mean=50, sd=10)
```



- This histogram of 1000 random data values from a normal population is a reasonable approximation of a normal distribution

Histogram of Data from Normal Population, $n = 100,000$

```
> Y <- rnorm(n=100000, mean=50, sd=10)
```



- This histogram of 100,000 random data values from a normal population is an excellent approximation of a normal curve, a better normal shape and smaller bins for a smoother curve

Normal Population and Corresponding Samples

Larger samples provide more information

- A distribution of *sample* values, even when from a normal population, never exactly conform to a perfect normal curve
- For small samples from a normal population, do not expect to see a shape even close to normality
- Only the distribution of values in very large samples, typically well beyond the sample size encountered in practice, closely approximate normality

3.2b
Standard Scores

Expression for the Normal Curve

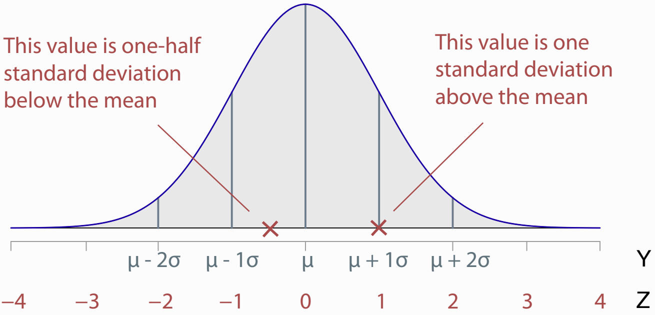
A mathematical abstraction

- The formula for a normal distribution, applied to the values of Y, generates a perfectly smooth bell-shaped curve described by f(Y), the height of the curve above each value of Y
- A *specific* normal curve also depends on the values of the population mean, μ, and the population standard deviation, σ

Formula for a normal curve: $$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$$

- π and e are mathematical constants, and μ and σ are constant for a specific curve, so, the only place Y appears in the expression is as a squared mean deviation score, $(Y-\mu)^2$
- **Key Concept**: This connection between the normal curve and the squared deviation score, and consequently the standard deviation, is central to statistical theory and analysis

Standard Deviation and the Normal Curve



For the normal distribution, the standard deviation becomes the natural scale for assessing how far a value is from the mean

z-value or standardized value: Number of standard deviations value Y_i is from its mean, regardless of the distribution, here illustrated for the normal distribution

z-value, An Example of Standardization

How many standard deviations is a value from its mean?

- To express the distance of the i^{th} data value from its mean in terms of standard deviations,

  population: $z_i = \dfrac{Y_i - \mu}{\sigma}$ sample: $z_i = \dfrac{Y_i - m}{s}$

- Standardized values are unitless, regardless if the original measures of Y are in dollars or kilograms, the corresponding z-values are expressed in terms of standard deviations
- **Key Concept**: Each individual value from *any* distribution for generic variable Y can be rescaled to a z-value, that is, standardized, providing two measurement scales, Y and Z
- **Standardized normal distribution**: A normal distribution expressed in the scale of standard scores, z-scores
- The concept of standardization applies to any distribution, but the standardized normal distribution is particularly useful

Illustration: Standard Scores

Compare test scores from two different tests

- Each of two groups of 18 newly hired employees were administered a different performance evaluation test, each test with a different number of items and standard deviation
  - Scores on the 1st test, Variable YA, ranged from 54 to a perfect score of 60, with $m = 56$ and $s = 1.782$
  - Scores on the 2nd test, Variable YB, ranged from 23 to a perfect score of 80, with the same $m = 56$, but $s = 16.606$
- Data:
  `https://web.pdx.edu/~gerbing/data/TestScores.csv`
- How can scores be compared across the two tests?

- Get standardized values, z-values, with the R `scale` function
- To work within the `d` data frame, create the variable of z-values, here called YA.z, with `lessR` function `Transform`

```
> d <- Transform(YA.z = scale(YA))
```

Illustration: Standard Scores

1^{st} set of test scores, Variable YA

- The first, and highest, score is $Y_1 = 60$, with a corresponding z-score of

$$z_1 = \frac{Y_1 - m}{s} = \frac{60 - 56}{1.782} = 2.24$$

- The test score of 60 is 2.24 standard deviations *above* the mean
- Similarly, the lowest score, Y_{18}, is 1.12 standard deviations *below* the mean
- Everyone did well in absolute scores, with scores ranging from 90% to 100%
- However, for the z-scores, the lowest scores of 90% correct were over a full standard deviation below the mean

	YA	YA.z
1	60	2.24
2	59	1.68
3	58	1.12
4	57	0.56
5	57	0.56
6	57	0.56
7	56	0.00
8	56	0.00
9	56	0.00
10	56	0.00
11	56	0.00
12	55	-0.56
13	55	-0.56
14	54	-1.12
15	54	-1.12
16	54	-1.12
17	54	-1.12
18	54	-1.12

Illustration: Standard Scores

2nd set of test scores, Variable YB

- These test scores are quite variable, with two people getting perfect scores of 80 and the lowest performer only achieving 23 out of 80 items, for 28.75%
- Even though the low scores were dramatically low, even the lowest score is less than two standard deviations below the mean
- The absolute scores in the two distributions exhibited two different patterns, one in which everyone did well vs one with considerably more variability
- Yet the ranges of *z*-scores are comparable in the two distributions

	YB	YB.z
1	80	1.45
2	80	1.45
3	78	1.32
4	74	1.08
5	69	0.78
6	65	0.54
7	62	0.36
8	60	0.24
9	55	-0.06
10	54	-0.12
11	52	-0.24
12	52	-0.24
13	47	-0.54
14	45	-0.66
15	43	-0.78
16	36	-1.20
17	33	-1.39
18	23	-1.99

Illustration: Conclusion

Absolute vs relative performance

- A value can be presented in its original units of measurement, Y_i, or, in terms of the standardized or z-value, Z_i
- **Absolute position**: Assessment of the position of one value in a distribution of values in terms of its magnitude, irrespective of the other values within the distribution
- Assess the absolute position with the original measurement, Y, or perhaps, if an evaluative test, expressed as the percentage correct
- **Relative position**: Assessment of the position of one value in a distribution of values compared to the position of the other values within the distribution
- **Key Concept**: The z-value indicates the *relative* position of the corresponding data value of variable Y within the distribution

3.2c
Normal Curve Probabilities

Concept of a Normal Curve Probability

Mathematical abstraction versus actual measurements

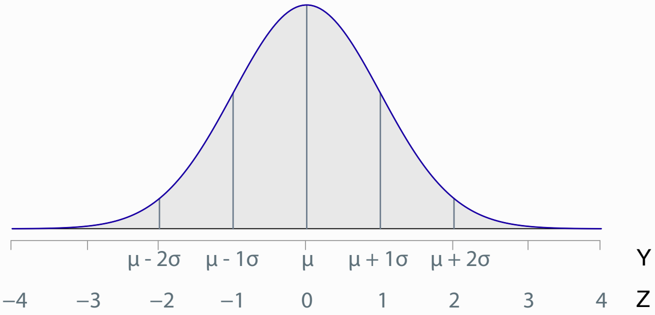
- The normal distribution is commonly encountered, both directly and indirectly, so an understanding of the probabilities that specific values occur is a task that underlies much data analysis
- A perfect normal distribution does *not* consist of actual measurements, but instead is defined as a mathematical abstraction of the values of a continuous variable
- The probability of any specific abstract value is zero because the width of any point on the real number line is 0
- **Key Concept**: A probability for a distribution defined on a continuous variable, such as the normal curve, only applies to an interval of values
- In graphic form, the probability corresponds to the area under the curve for the specified interval

Normal Curve Probability

Definition and examples

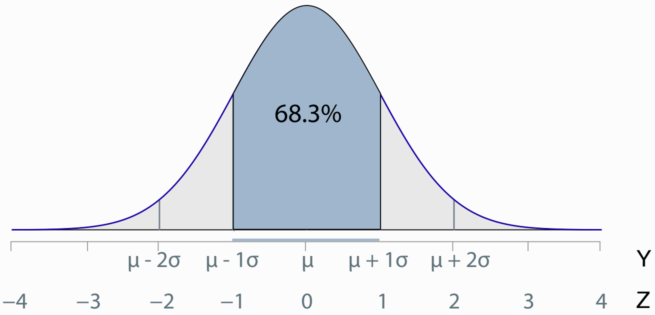
- **Normal curve probability**: The probability that a randomly selected value from a normal distribution is within a specified range of values
  - Ex: Probability, for the normal distribution in which $\mu = 50$ and $\sigma = 10$, that a randomly selected value is between 50 and 60, that is, $P(50 < Y < 60)$
  - Ex: Probability that an applicant's score on the GMAT is larger than the informal cutoff of 650 used by many top graduate programs, or, $P(Y > 650)$, relative to an approximately normal distribution of GMAT values with $\mu = 525$ and $\sigma = 100$
- **Key Concept**: Normal curve probabilities are directly related to the standard deviation of the specific normal curve of interest and the associated z-scores

Standard Deviation, Probability and the Normal Curve



For a normal distribution, a standardized value, a *z*-value, relates to the probability of the occurrence of a value within a given range of values

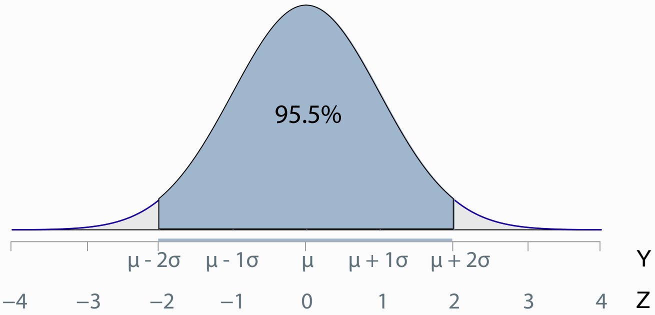
Standard Deviation, Probability and the Normal Curve



More than 68% of all values in a normal distribution, more than 2/3, fall within 1 standard deviation about the mean

So over 68% of all values in a standardized normal distribution of *z*-values fall between -1 and 1

Standard Deviation, Probability and the Normal Curve



More than 95% of all values in a normal distribution fall within 2 standard deviations about the mean

So over 95% of all values in a standardized normal distribution of *z*-values fall between -2 and 2

Normal Curve Probability

- Use the lessR `prob_norm()` function to obtain a normal curve probability and the accompanying graph

  `> prob_norm(lo=1, hi=2)`, with default $\mu = 0$, $\sigma = 1$

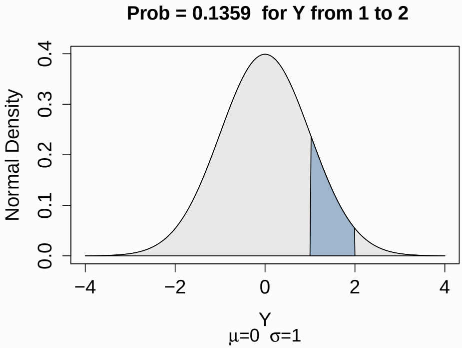


Figure: Illustrated probability of a randomly sampled value between 1 and 2 for the normal distribution with $\mu = 0$ and $\sigma = 1$

Normal Curve Tail Probability

- For a normal curve tail probability with `prob_norm()`, specify just a `lo` or `hi` value, here for $P(Y > 650)$ on the GMAT

```
> prob_norm(lo=650, mu=525, sigma=100)
```

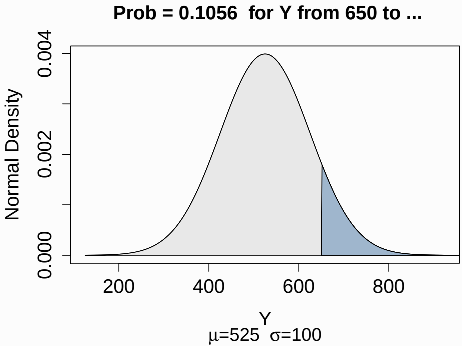


Figure: Illustrated tail probability that about 10.5% of GMAT total scores are greater than 650, with $\mu = 525$ and $\sigma = 100$

The Normal Curve: Tail Probabilities

"Normal" means close to the middle

range of z-values	% values WITHIN this range	% values OUTSIDE this range
-1 and 1	68.2689492137%	31.7310507863%
-2 and 2	95.4499736104%	04.5500263896%
-3 and 3	99.7300203937%	00.2699796063%
-4 and 4	99.9936657516%	00.0063342484%
-5 and 5	99.9999426697%	00.0000573303%
-6 and 6	99.9999998027%	00.0000001973%
-7 and 7	99.9999999997%	00.0000000003%

The Normal Curve: Tail Probabilities

"Normal" means close to the middle

- Almost 1/3 of normally distributed values are outside of 1 standard deviation around the mean
- Less than 5% of normally distributed values are further than 2 standard deviations from the mean
- Less than 0.27% of normally distributed values are further than 3 standard deviations from the mean, becoming rare
- Less than 1 per ten thousand of normally distributed values are further than 4 standard deviations from the mean
- Less than 1 per million of normally distributed values are further than 5 standard deviations from the mean
- Less than 2 per billion of normally distributed values are further than 6 standard deviations from the mean
- Less than 3 per trillion of normally distributed values are further than 7 standard deviations from the mean

Normal Distribution Quantiles

What values about μ contain 95% of the distribution?

- k^{th} **Quantile**: Value greater than k% of distribution
- So k^{th} quantile also cuts off $1 - k\%$ *above* the value
- $z_{.025}$ refers to the quantile that cuts off the top 2.5% of the distribution

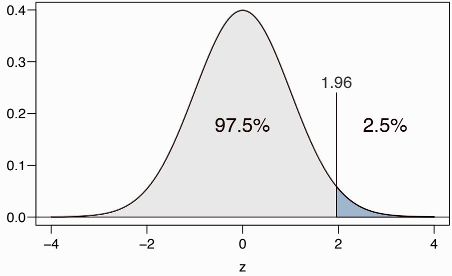


Figure: The .975 quantile, $z_{.025} = 1.96$, cuts off bottom 97.5% of the standard normal distribution, and top 2.5%.

- **Key Concept**: 95% of the values from a normal distribution of Y are within 1.96 standard deviations of the mean of Y, that is, between $z_{.025} = 1.96$ and $-z_{.025} = -1.96$

3.2d
General Density Curves

Move Beyond the Arbitrary Histogram

Histograms are pre-computer technology

- A histogram is the traditional analysis since the 19^{th} century for graphing the distribution of a continuous variable
- Unfortunately, as we have seen, a histogram suffers from two artifacts: bin width and bin shift
- An even more basic issue is that a histogram groups data into bins, yet the distribution that characterizes the values of a continuous variable, such as the normal curve, is realized by a continuous variable that graphs as a smooth curve
- The underlying smooth curve, in many situations, is a normal curve, but many other possibilities also exist
- **Key Concept**: With modern software, move beyond the artifacts and approximations of a histogram by also obtaining the estimated underlying smooth distribution curve

Smooth Curve Estimated from the Data

- A plot of smooth, idealized distribution, a smoothed-out histogram, is called a **density curve**
- **Qualitative interpretation** of density: Identify the overall shape of the underlying continuous distribution
- For example, for a normal distribution, identify
  - The mode, the value with the largest density, which for a small range of values about that value, corresponds to the most frequently occurring values
  - The tails, which contain the values that rarely occur

Qualitative Interpretation of Densities

Understand the general characteristics of the distribution

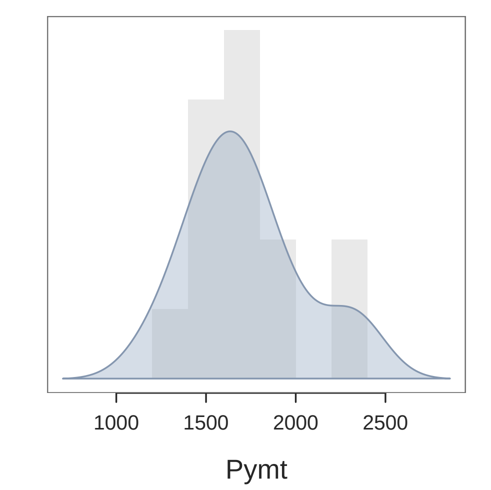
- Consider again the distribution of Pymt, the Monthly Mortgage Payment, and the plot of the corresponding 14 data values

```
> d <-
   Read("https:
//web.pdx.edu/~gerbing/data/mortgage.csv")
```

- To plot the smoothed, density curve, use the `lessR` parameter `density` set to `TRUE`.
- By default, the plot is of the estimated generalized smooth curve. Add `type="both"` to also plot the estimated normal curve, both superimposed over a histogram of the data.

Smooth Curve Estimated from the Data

```
> X(Pymt, type="density")
```



Density Function Options

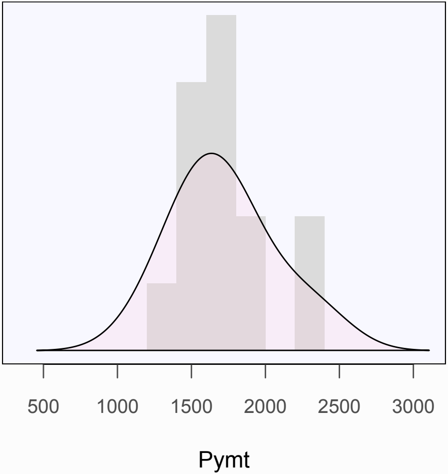
Obtain explicit control of aspects of the graph

- The `X()` options `bin_start` and `bin_width` to specify the bins of the histogram also apply to `type="density"`
- Control the smoothness of the generalized curve with the bandwidth parameter, bw
- From the output:

```
Density bandwidth for general curve: 138.5702
For a smoother curve, increase bandwidth with option: bw
```

Smooth Curve Estimated from the Data with Options

```
> X(Pymt, bw=250, type="density")
```



Index

Subtract 2 from each listed value to get the Slide #

- The End