Data Exploration

- Getting To Know The Data
- Handling Data Quality Issues
- Visualizing Relationships Between Features
- Data Preparation

The Data Quality Report

- Tabular reports that describe the characteristics of each feature using statistical **central tendency** and **variation**.
- Accompanied by data visualizations: A **histogram** for each continuous feature and a **bar plot** for each categorical feature.

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___

Table: Portions of the ABT for the motor insurance claims fraud detection problem.

ID	Type	Inc.	Marital Status	Num Clmnts.	Injury Type	Hospital Stay	Claim Amnt.	Total Claimed	Num Claims	Num Soft Tiss.	% Soft Tiss.	Claim Amt Rcvd.	Fraud Flag
1	CI	0		2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1		0	15,028	0
3	CI	54,613	Married	1	Broken Limb	No	-99,999	0	0	0	0	572	0
4	CI	0		4	Broken Limb	Yes	5,097	11,661	1	1	1.0	7,864	0
5	CI	0		4	Soft Tissue	No	8869	0	0	0	0	0	1
6	CI	0		1	Broken Limb	Yes	17,480	0	0	0	0	17,480	0
7	CI	52,567	Single	3	Broken Limb	No	3,017	18,102	2	1	0.5	0	1
8	CI	0		2	Back	Yes	7463	0	0	0	0	7,463	0
9	CI	0		1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300	Married	4	Back	No	2,260	0	0	0	0	2,260	0
		⋮				⋮					⋮		
300	CI	0		2	Broken Limb	No	2,244	0	0	0	0	2,244	0
301	CI	0		1	Broken Limb	No	1,627	92,283	3	0	0	1,627	0
302	CI	0		3	Serious	Yes	270,200	0	0	0	0	270,200	0
303	CI	0		1	Soft Tissue	No	7,668	92,806	3	0	0	7,668	0
304	CI	46,365	Married	1	Back	No	3,217	0	0		0	1,653	0
		⋮				⋮					⋮		
458	CI	48,176	Married	3	Soft Tissue	Yes	4,653	8,203	1	0	0	4,653	0
459	CI	0		1	Soft Tissue	Yes	881	51,245	3	0	0	0	1
460	CI	0		3	Back	No	8,688	729,792	56	5	0.08	8,688	0
461	CI	47,371	Divorced	1	Broken Limb	Yes	3,194	11,668	1	0	0	3,194	0
462	CI	0		1	Soft Tissue	No	6,821	0	0	0	0	0	1
		⋮				⋮					⋮		
491	CI	40,204	Single	1	Back	No	75,748	11,116	1	0	0	0	1
492	CI	0		1	Broken Limb	No	6,172	6,041	1		0	6,172	0
493	CI	0		1	Soft Tissue	Yes	2,569	20,055	1	0	0	2,569	0
494	CI	31,951	Married	1	Broken Limb	No	5,227	22,095	1	0	0	5,227	0
495	CI	0		2	Back	No	3,813	9,882	3	0	0	0	1
496	CI	0		1	Soft Tissue	No	2,118	0	0	0	0	0	1
497	CI	29,280	Married	4	Broken Limb	Yes	3,199	0	0	0	0	0	1
498	CI	0		1	Broken Limb	Yes	32,469	0	0	0	0	16,763	0
499	CI	46,683	Married	1	Broken Limb	No	179,448	0	0		0	179,448	0
500	CI	0		1	Broken Limb	No	8,259	0	0	0	0	0	1

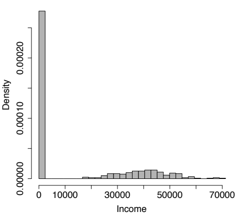
Table: A data quality report for the motor insurance claims fraud detection ABT

(a) Continuous Features

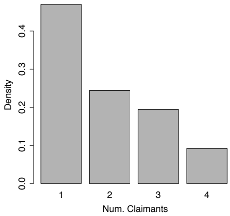
Feature	Count	% Miss.	Card.	Min	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(b) Categorical Features

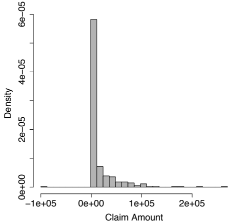
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
INSURANCE TYPE	500	0.0	1	CI	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	Married	99	51.0	Single	48	24.7
INJURY TYPE	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
HOSPITAL STAY	500	0.0	2	No	354	70.8	Yes	146	29.2



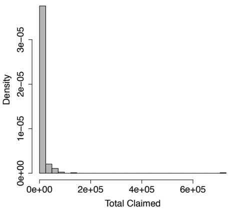
(a) INCOME



(b) NUM CLAIMANTS

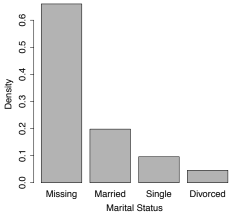


(c) CLAIM AMOUNT

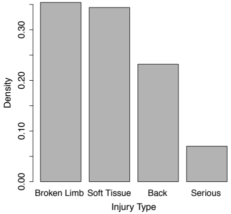


(d) TOTAL CLAIMED

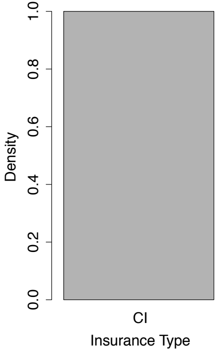
Visualizations of the continuous and categorical features



(e) MARITAL STATUS



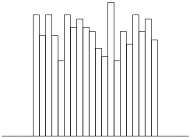
(f) INJURY TYPE



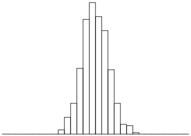
(g) INSURANCE TYPE

Visualizations of the continuous and categorical features

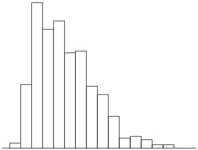
- For categorical features, we should:
  - Examine the mode, 2^{nd} mode, mode %, and 2^{nd} mode % as these tell us the most common levels within these features and will identify if any levels dominate the dataset.
- For continuous features we should:
  - Examine the mean and standard deviation of each feature to get a sense of the central tendency and variation of the values within the dataset for the feature.
  - Examine the minimum and maximum values to understand the range that is possible for each feature.



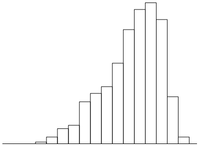
(h) Uniform



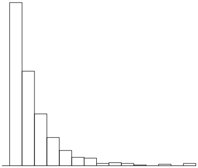
(i) Normal (Unimodal)



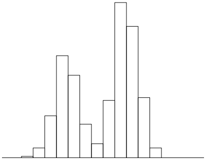
(j) Unimodal (skewed right)



(k) Unimodal (skewed left)



(l) Exponential



(m) Multimodal

- The probability density function for the **normal** distribution (or **Gaussian distribution**) is

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

where x is any value, and μ and σ are parameters that define the shape of the distribution: the **population mean** and **population standard deviation**.

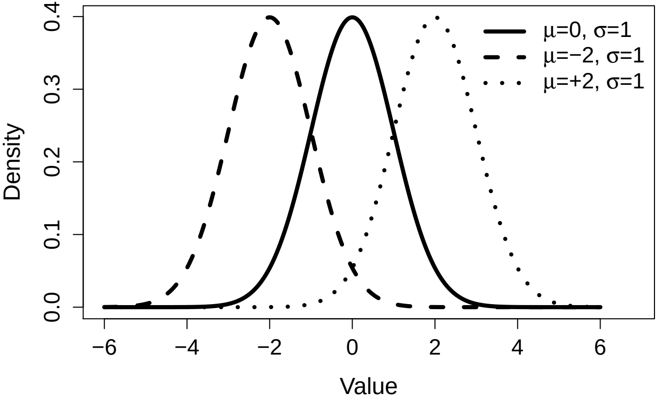


Figure: Three normal distributions with different means but identical standard deviations.

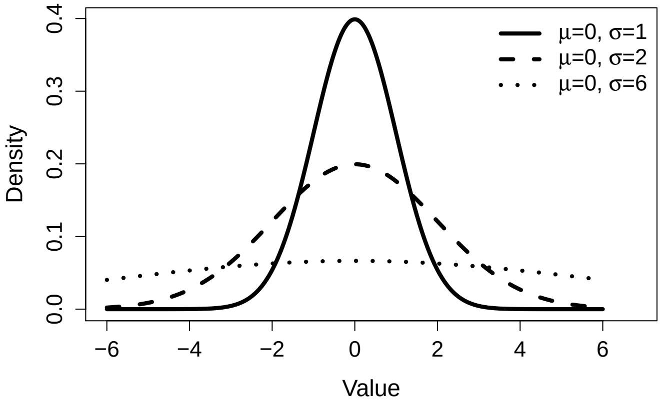


Figure: Three normal distributions with identical means but different standard deviations.

- The $68 - 95 - 99.7$ rule is a useful characteristic of the normal distribution.
- The rule states that approximately:
  - 68% of the observations will be within one σ of μ
  - 95% of observations will be within two σ of μ
  - 99.7% of observations will be within three σ of μ.

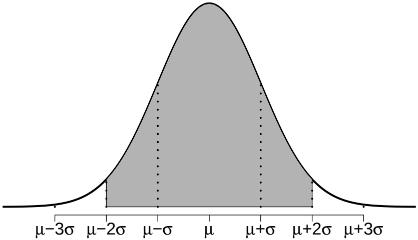


Figure: An illustration of the $68 - 95 - 99.7$ percentage rule that a normal distribution defines as the expected distribution of observations. The grey region defines the area where 95% of observations are expected.

Case Study: Motor Insurance Fraud

Examine the data quality report for the motor insurance fraud prediction scenario and comment on the data quality issues.

- A **data quality issue** is loosely defined as anything *unusual* about the data.
- The most common data quality issues are:
  - **missing values**
  - **outliers**

The data quality plan for the motor insurance fraud prediction.

Feature	Data Quality Issue	Potential Handling Strategies
NUM SOFT TISSUE	Missing values (2%)	
CLAIM AMOUNT	Outliers (high)	
AMOUNT RECEIVED	Outliers (high)	

Handling Missing Values

- Approach 1: Drop any features that have missing value.
- Approach 2: Apply **complete case analysis**.
- Approach 3: Derive a **missing indicator feature** from features with missing value.

- **Imputation** replaces missing feature values with a plausible estimated value based on the feature values that are present.
- The most common approach to imputation is to replace missing values for a feature with a measure of the central tendency of that feature.
- We would be reluctant to use imputation on features missing in excess of 30% of their values and would strongly recommend against the use of imputation on features missing in excess of 50% of their values.

Handling Outliers

- The easiest way to handle outliers is to use a **clamp transformation** that clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the offending outliers

$$a_i = \begin{cases} lower & \text{if } a_i < lower \\ upper & \text{if } a_i > upper \\ a_i & otherwise \end{cases} \quad (2)$$

where a_i is a specific value of feature a, and *lower* and *upper* are the lower and upper thresholds.

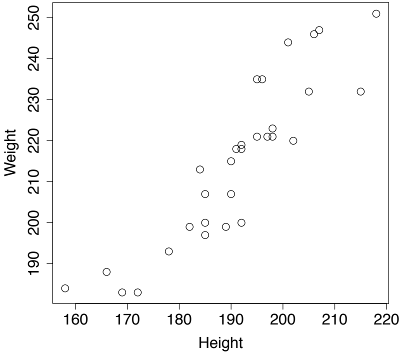
Case Study: Motor Insurance Fraud

Table: The data quality plan for the motor insurance fraud prediction ABT.

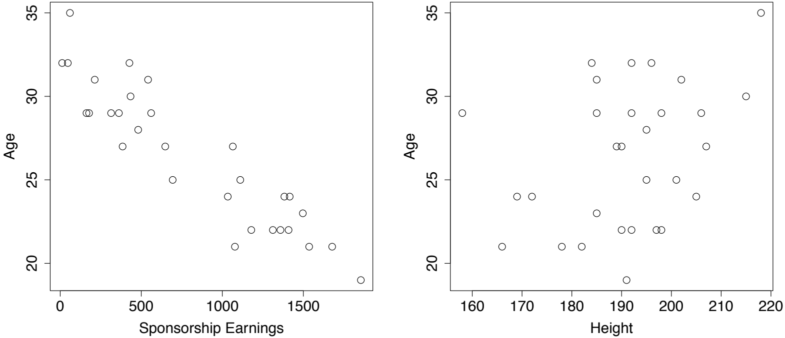
Feature	Data Quality Issue	Potential Handling Strategies
NUM SOFT TISSUE	Missing values (2%)	Imputation (median: 0.0)
CLAIM AMOUNT	Outliers (high)	Clamp transformation (manual: 0, 80 000)
AMOUNT RECEIVED	Outliers (high)	Clamp transformation (manual: 0, 80 000)

ID	Position	Height	Weight	Career Stage	Age	Sponsorship Earnings	Shoe Sponsor
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

A **scatter plot** is based on two axes: the horizontal axis represents one feature and the vertical axis represents a second.



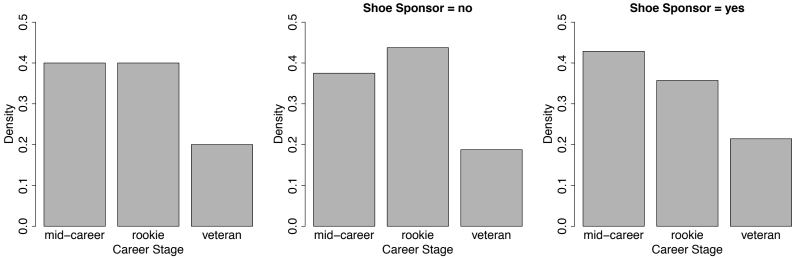
An example showing the relationship between the HEIGHT and WEIGHT features from the professional basketball squad dataset



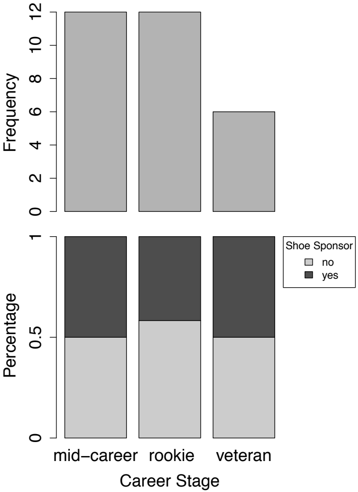
(a) (b)

(a) the strong negative covariance between the SPONSORSHIP EARNINGS and AGE features and (b) the HEIGHT and AGE features from the dataset

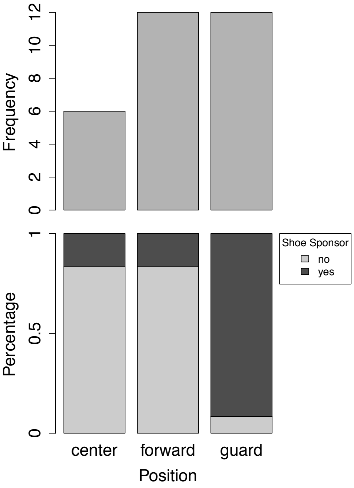
Visualize the relationship between categorical variables with small multiple bar plots



Using small multiple bar plot visualizations to illustrate the relationship between the CAREER STAGE and SHOE SPONSOR features.



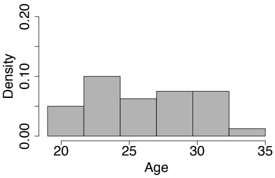
(c) Career Stage and Shoe Sponsor



(d) Position and Shoe Sponsor

If the number of levels of one of the features being compared is no more than three we can use **stacked bar plots**

- To visualize the relationship between a continuous feature and a categorical feature a **small multiples** approach that draws a histogram of the values of the continuous feature for each level of the categorical feature is useful.

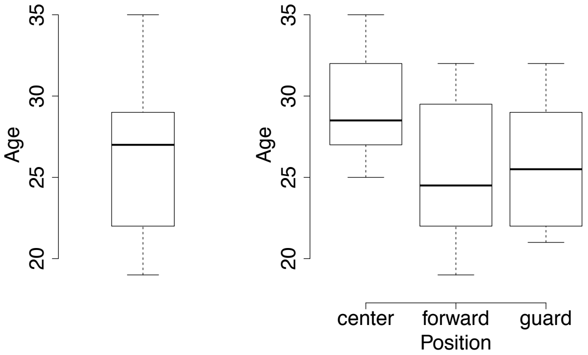


(e) Age

(f) Age and Position

Using small multiple histograms to visualize the relationship between the AGE feature and the POSITION FEATURE.

- A second approach to visualizing the relationship between a categorical feature and a continuous feature is to use a collection of box plots.
- For each level of the categorical feature a box plot of the corresponding values of the continuous feature is drawn.



(g) Age (h) Age and Position

Using box plots to visualize the relationship between the AGE and the POSITION feature.

Data Preparation

- Some data preparation techniques change the way data is represented just to make it more compatible with certain machine learning algorithms.
    - Normalization
    - Binning
    - Sampling

Normalization

- **Normalization** techniques change a continuous feature to fall within a specified range while maintaining the relative differences between the values for the feature.
- **Range normalization** converts a feature value into the range [*low*, *high*] as

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low \tag{3}$$

- **Standard score** measures how many standard deviations a feature value is from the mean for that feature.

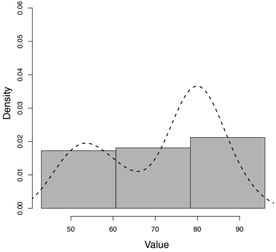
$$a_i' = \frac{a_i - \bar{a}}{sd(a)} \tag{4}$$

The result of normalising a small sample of the HEIGHT and SPONSORSHIP EARNINGS features from the professional basketball squad dataset.

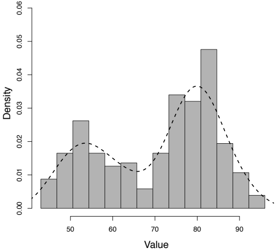
	HEIGHT			SPONSORSHIP EARNINGS		
	Values	Range	Standard	Values	Range	Standard
	192	0.500	-0.073	561	0.315	-0.649
	197	0.679	0.533	1,312	0.776	0.762
	192	0.500	-0.073	1,359	0.804	0.850
	182	0.143	-1.283	1,678	1.000	1.449
	206	1.000	1.622	314	0.164	-1.114
	192	0.500	-0.073	427	0.233	-0.901
	190	0.429	-0.315	1,179	0.694	0.512
	178	0.000	-1.767	1,078	0.632	0.322
	196	0.643	0.412	47	0.000	-1.615
	201	0.821	1.017	1111	0.652	0.384
Max	206			1,678		
Min	178			47		
Mean	193			907		
Std Dev	8.26			532.18		

Binning

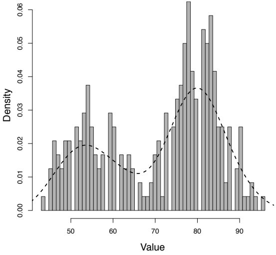
- **Binning** involves converting a continuous feature into a categorical feature.
- To perform binning, we define a series of ranges (called **bins**) for the continuous feature that correspond to the levels of the new categorical feature we are creating.
- Deciding on the number of bins can be difficult. The general trade-off is this:
  - If we set the number of bins to a very low number we may lose a lot of information
  - If we set the number of bins to a very high number then we might have very few instances in each bin or even end up with empty bins.



(i) 3 bins

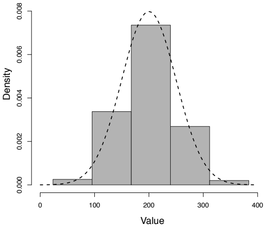


(j) 14 bins

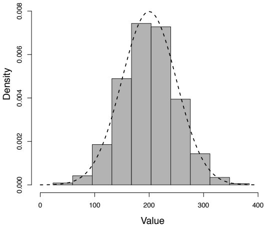


(k) 60 bins

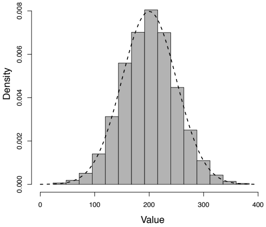
- The **equal-width binning** splits the range of the feature values into b bins each of size $\frac{range}{b}$.
- **Equal-frequency binning** first sorts the continuous feature values into ascending order and then places an equal number of instances into each bin, starting with bin 1.
  - The number of instances placed in each bin is simply the total number of instances divided by the number of bins, b.



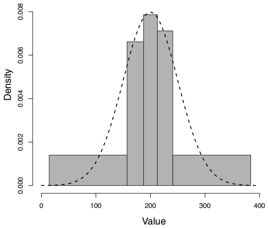
(l) 5 Equal-width bins



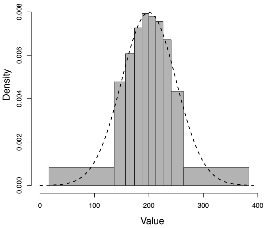
(m) 10 Equal-width bins



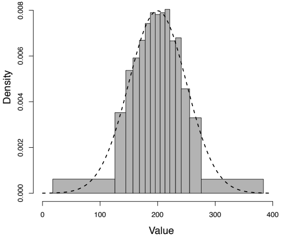
(n) 15 Equal-width bins



(o) 5 Equal-frequency bins



(p) 10 Equal-frequency bins



(q) 15 Equal-frequency bins

Sampling

- Sometimes the dataset we have is so large that we do not use all the data available to us in an ABT and instead **sample** a smaller percentage from the larger dataset.
- We need to be careful when sampling, however, to ensure that the resulting datasets are still representative of the original data and that no unintended **bias** is introduced during this process.
- Common forms of sampling include:
  - **top sampling**
  - **random sampling**
  - **stratified sampling**
  - **under-sampling**
  - **over-sampling**

- **Stratified sampling** is a sampling method that ensures that the relative frequencies of the levels of a specific **stratification feature** are maintained in the sampled dataset.
- To perform stratified sampling:
  - the instances in a dataset are divided into groups (or strata), where each group contains only instances that have a particular level for the stratification feature
  - $s\%$ of the instances in each stratum are randomly selected
  - these selections are combined to give an overall sample of $s\%$ of the original dataset.

- **Under-sampling** begins by dividing a dataset into groups, where each group contains only instances that have a particular level for the feature to be under-sampled.
- The number of instances in the *smallest* group is the under-sampling target size.
- Each group containing more instances than the smallest one is then randomly sampled by the appropriate percentage to create a subset that is the under-sampling target size.
- These under-sampled groups are then combined to create the overall under-sampled dataset.

- **Over-sampling** addresses the same issue as under-sampling but in the opposite way around.
- After dividing the dataset into groups, the number of instances in the *largest* group becomes the over-sampling target size.
- From each smaller group, we then create a sample containing that number of instances using **random sampling with replacement**.
- These larger samples are combined to form the overall over-sampled dataset.