Statistical Analysis

COSC 757 Spring 2016

Statistical Inference

- A widespread tool for performing estimation and prediction is *statistical inference*.
- Statistical Inference
  - Methods for estimating and testing hypotheses about *population* characteristics based on information contained in a sample
- Population
  - A population is a collection of <u>all</u> elements of interest for a particular study
  - A *parameter* is a characteristic of a population, such as the mean number of customer service calls of all cell phone customers
    - Example: Cell phone company wants actionable results for all their present and future customers (population), not only the 3333 customers for which they gathered the data (sample)
- Sample
  - A sample is a <u>representative subset</u> of the population
  - If the sample characteristics deviate systematically from the population characteristics, *statistical inference should not be applied*
  - A *statistic* is a characteristic of a sample, such as the mean number of customer service calls of the 3333 customers in the sample (1.563).

Statistical Inference (*cont'd*)

- The values of population parameters are *unknown* for most problems
  - Specifically, the value of the population mean is usually unknown
  - Example: We do not know the true mean number (μ) of customer service calls to be made by all of the company's cell phone customers
  - To represent their unknown nature, population parameters are often denoted with Greek letters
  - Instead, data analysts use estimation
    - For example, estimate the unknown value of population mean (μ) from the computed sample mean $(\bar{x})$
    - Example: We would estimate the mean number of customer service calls for all customers to be 1.563, since this is the value of our observed sample mean
    - Caveat: Estimation only valid if sample is representative of population
      - Example: This would not be possible if the sample dataset included only disgruntled customers!

	Sample Statistics	…Estimates…	Population Parameters
Mean	$\bar{x}$	→	μ
Standard deviation	s	→	σ
Proportion	p	→	π

Statistical Inference (*cont'd*)

- Proportion
  - The sample proportion *p*, is the statistic used to measure the unknown value of the population proportion π.
  - Example: In Chapter 3, we found proportion of churners in dataset is *p*=0.145, which could be used to estimate true proportion of churners in the population of all customers
- Point estimation
  - Use of a single known value of a statistic to estimate the associated population parameter
  - The observed value of the statistic is called the *point estimate*
- Any statistic observed from sample data may be used to estimate the analogous parameter in the population
- *Any* sample characteristic is a statistic, which could be used to estimate its respective parameter

Statistical Inference (*cont'd*)

- Example
  - The proportion of those in the population who "churn" is unknown
  - The proportion of those in the population who churn is estimated from the proportion of those in the sample who churn
  - The sample proportion $p = 0.145$ is used to estimate the population proportion who churn
  - In these examples, our estimation is only valid when the sample is truly representative of the entire population

How Confident Are We in Our Estimates?

- Is the population number of customer service calls the same as the sample mean $\bar{x}$=1.563? Probably not.
  - Population contains more information than the sample
  - Our point estimates will nearly always “miss” the target parameter by a certain amount
- Sampling Error
  - Distance between the observed value of the point estimate and the unknown value of its target parameter
  - Defined as: $|statistic\ - parameter|$
  - For example, the sampling error for the mean is: (always positive)

$$|\bar{x} - \mu|$$

  - Value of the sampling error is usually unknown
    - For continuous variables, the probability that the observed value of a point estimate exactly equals its target parameter is precisely zero

How Confident Are We in Our Estimates? (*cont'd*)

- Point estimates have <u>no measure of confidence</u> in their accuracy
  - There is no probability statement associated with the estimate
  - Estimate is <u>probably</u> close to the value of the target parameter, but <u>possibly it may be far</u> away
  - Point estimation can be compared to throwing darts with infinitesimally small tips to a vanishingly small bull's-eye
    - Worse, the bull's-eye is hidden, and the thrower will never know for sure how close the darts are to the target
  - As an alternative the dart thrower could throw a wide beer mug to the target
    - There does indeed exist a positive probability that some portion of the mug has hit the hidden bull's-eye
    - While not sure, we can have a certain degree of confidence that the target has been hit
    - Very roughly, the beer mug represents our next estimation method, *confidence intervals*

Confidence Interval Estimation

- Confidence Interval Estimate
  - Consists of an interval of numbers produced by a point estimate
  - Includes a <u>confidence level</u> specifying the probability the interval contains the population parameter
  - Confidence intervals have the general form:

    Point Estimate +/- Margin of Error
  - The margin of error is a measure of precision, and smaller values indicate greater precision

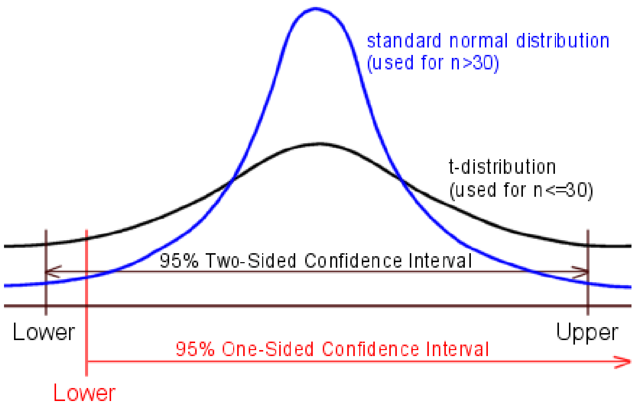
Confidence Interval Estimation (*cont'd*)

- Example
  - The *t*-interval for the population mean is:

$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$$

$$\text{Where, the Point Estimate} = \bar{x}, \text{ and Margin of Error} = t_{\alpha/2}(s/\sqrt{n})$$

  - The *t*-interval for the mean is used when either: the population has a normal distribution; or the sample size *n* is large
- For normal distribution (n>30)
  - $t_{\alpha/2}$ = 1.645 for 90% confidence
  - $t_{\alpha/2}$ = 1.960 for 95% confidence
  - $t_{\alpha/2}$ = 2.576 for 99% confidence



Confidence Interval Estimation (*cont'd*)

- The standard error of the sample mean is represented by:

$$(s/\sqrt{n})$$

- Therefore, the standard error is minimized whenever the sample size *n* is large, or the sample variability is small
- In addition, the value for the multiplier ($t_{\alpha/2}$) is associated with the sample size and confidence level (usually 90% – 99%) specified by the analyst and is smaller for lower confidence levels
- A more precise confidence interval estimate is achieved by increasing the sample size *n*
- Typically, finding a large sample size is not a particular problem in data mining

Confidence Interval Estimation
(*cont'd*)

- Example
  - Find the 95% confidence *t*-interval for the mean number of Customer Service Calls
  - Recall that the sample mean = 1.563, and sample standard deviation = 1.315

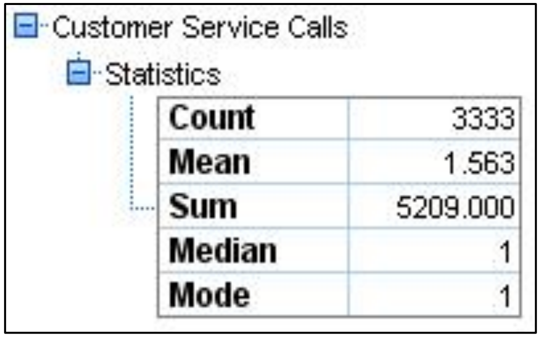
$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$$

$$1.563 \pm 1.961(1.315/\sqrt{3333})$$

$$1.563 \pm 0.045$$

$$(1.518, 1.608)$$

  - We are ~95% confident the population mean number of Customer Service Calls is between 1.518 and 1.608. Our margin of error is 0.045


Customer Service Calls
Statistics

Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

Confidence Interval Estimation
(*cont'd*)

- Data miners are often called upon to perform *subgroup analyses*
  - This is, to estimate the behavior of specific subsets of customers instead of the entire customer base
  - Example: Estimate the mean number of customer service calls for customers who have both the International Plan and the VoiceMail Plan and who have more than 220 day minutes
  - We are 95% confident the population mean number of Customer Service Calls is between 0.875 and 2.339. Our margin of error is 0.732
  - Here, our estimate for this specific subset of customers is <u>less precise</u> than our estimate obtained for the entire customer base in the previous Example

Customer Service Calls

Statistics

Count	28
Mean	1.607
Standard Deviation	1.892

$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$$

$$1.607 \pm 2.048(1.892/\sqrt{28})$$

$$1.607 \pm 0.732$$

$$(0.875, 2.339)$$

Confidence Interval Estimation
(*cont'd*)

- Conclusion
  - Confidence interval estimation can be applied to any desired target parameter
  - Most widely used in statistical practice to estimate the population mean, population standard deviation, and population proportion of successes

How to reduce the margin of error

- The margin of error *E* for a 95% confidence interval for the population mean μ can be interpreted as:

  "We can estimate μ to within E units with 95% confidence."

- The smaller the margin of error, the more precise our estimation
- *How can we reduce our margin of error*?
- Margin of error *E* contains three quantities
  - $t_{\alpha/2}$, which depends on the confidence level
  - s, which is a characteristic of the data, and <u>may not be changed</u>
  - n, the sample size
- Thus, we can decrease the margin of error (E) by
  - Decreasing the confidence level, $t_{\alpha/2}$ - NOT RECOMMENDED
  - Increasing sample size, n - RECOMMENDED
- Increasing the sample size is the only way to decrease the margin of error while maintaining a constant level of confidence

How to reduce the margin of error (*cont'd*)

- Example: We have a new sample with 5000 customers
  - Calculate the new margin of error for the 95% interval of the mean number of customer service calls

$$E = t_{\alpha/2}(s/\sqrt{n})$$
$$E = 1.96(1.315/\sqrt{5000})$$
$$E = 0.036$$

  - Due to the $\sqrt{n}$ in the formula for E, an increase of a in the sample size leads to a reduction of margin of error of $\sqrt{a}$

Customer Service Calls

Statistics

Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

Confidence Interval Estimation of the Proportion

- Example: 483 of 3333 customers churned
- An estimate of the *population proportion* π of all company's customers who churn is:

$$p = \frac{number\ who\ churn}{sample\ size} = \frac{x}{n} = \frac{483}{3333} = 0.1449$$

- But we have no measure of confidence on the accuracy of this estimate
- A *confidence interval for the population proportion* π is

$$p \pm Z_{\propto/2}\sqrt{\frac{p \cdot (1-p)}{n}}$$

- Where the sampling proportion p is the point estimate of π and $Z_{\propto/2}\sqrt{\frac{p \cdot (1-p)}{n}}$ represents the margin of error. $Z_{\propto/2}$ represents the confidence level (1.645 for 90%, 1.96 for 95% and 2.576 for 99%)
- The Z-interval for π may be used whenever both $np > 5$ and $n(1-p) > 5$

Confidence Interval Estimation of the Proportion (*cont'd*)

- Example: Calculate 95% confidence interval for the proportion π of churners among the entire population of the company's customers

$$p \pm Z_{\propto/2}\sqrt{\frac{p\cdot(1-p)}{n}}$$
$$= 0.1149 \pm 1.96\sqrt{\frac{(0.1449)(0.8551)}{333}}$$
$$= 0.1149 \pm 0.012$$
$$= (0.1329, 0.1569)$$

- We are 95% confident that this interval captures the population proportion π. The confidence interval for π takes the form

$$p \pm E = 0.1149 \pm 0.012$$

- The margin of error $E = Z_{\propto/2}\sqrt{\frac{p\cdot(1-p)}{n}}$ can be interpreted as follows:

"We can estimate π to within E with 95% of confidence"

- In this case, we estimate proportion of churners to within 0.012 (or 1.2%) with 95% confidence.
- Again, given a confidence level, margin of error can be reduced by using a larger sample

Hypothesis Testing for the Mean

- Hypothesis testing evaluates claims about a population parameter using evidence from the sample
- Two competing statements (hypotheses) are crafted about the parameter value:
  - *Null Hypothesis* (H_0) – represents assumed value
  - *Alternative Hypothesis* (H_a) – represents alternative claim about the value
- There are two possible conclusions:
  - Reject H_0
  - Do not reject H_0

Hypothesis Testing for the Mean (*cont'd*)

- Example: A criminal trial is a form of hypothesis test, with following hypotheses:
  - H_0: Defendant is innocent
  - H_a: Defendant is guilty
  - Type I error: Reject H_0 when H_0 is true – Jury convicts innocent person
  - Type II error: Do not reject H_0 when H_0 is false – Jury acquits a guilty person
  - Correct: Reject H_0 when H_0 is false – Jury convicts a guilty person
  - Correct: Do not reject H_0 when H_0 is false – Jury acquits an innocent person

		Reality	
		H_0 true: Defendant did not commit crime	H_0 false: Defendant did commit crime
Jury's decision	Reject H_0: Find defendant guilty	Type I error	Correct decision
	Do not reject H_0: Find defendant not guilty	Correct decision	Type II error

Hypothesis Testing for the Mean (*cont'd*)

- Probability of Type I error is denoted α
- Probability of Type II error is denoted β
  - For constant sample size, decrease in α is associated with increase in β and vice versa
- Common to restrict the hypotheses to these three forms
  - Left-tailed test; $H_0\text{: } \mu \geq \mu_0 \ \ vs.\ H_a\text{: } \mu < \mu_0$
  - Right-railed test; $H_0\text{: } \mu \leq \mu_0 \ \ vs.\ H_a\text{: } \mu > \mu_0$
  - Two-tailed test; $H_0\text{: } \mu = \mu_0 \ \ vs.\ H_a\text{: } \mu \neq \mu_0$

  where μ_0 represents a hypothesized value of μ

Hypothesis Testing for the Mean (*cont'd*)

- When sample size is large or population is normally distributed, the test statistic

$$t_{data} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

follows a *t* distribution, with $1 - n$ degrees of freedom.

- The value of t_{data} is interpreted as the number of standard errors above/below the hypothesized mean μ that the sample mean $\bar{x}$ resides, where the standard error equals $s/\sqrt{n}$
- When the value of t_{data} is extreme, this indicates conflict between null hypothesis (with μ_0 and the observed data) and the observed data
- When t_{data} is extreme, the null hypothesis H_0 is rejected
- How extreme is extreme? This is measured using the *p-value*

Hypothesis Testing for the Mean (*cont'd*)

- The *p-value* is the probability of observing a sample statistic as extreme as the statistic actually observed
- *p-value* is a probability, and its value falls between 0 and 1
- The names of the forms of the hypothesis test indicate in which tail or tails of the t distribution the p-value will be found
- We will reject H_0 if the p-value is small
- Researchers set the level of significance α at some small value (such as 0.05); p-value is small if it is less than α:

Reject H_0 if the p-value is $< \alpha$

Form of Hypothesis Test	**p-Value**
Left-tailed test. H_0: $\mu \geq \mu_0$ $vs.$ H_a: $\mu < \mu_0$	$P(t < t_{data})$
Right-tailed test. H_0: $\mu \leq \mu_0$ $vs.$ H_a: $\mu > \mu_0$	$P(t > t_{data})$
Two-tailed test. H_0: $\mu = \mu_0$ $vs.$ H_a: $\mu \neq \mu_0$	If $t_{data} < 0$ then p-value = $2 \cdot P(t < t_{data})$. If $t_{data} > 0$ then p-value = $2 \cdot P(t > t_{data})$.

Hypothesis Testing for the Mean (*cont'd*)

- Example: Test whether mean number of Customer Service Calls for customers with International Plan and VoiceMail Plan and who have more than 220 day minutes differs from 2.4, with a level of significance α=0.05

$$H_0\text{: } \mu = 2.4 \;\; vs.\; H_a\text{: } \mu \neq 2.4$$

- Null hypothesis is rejected if p-value<0.05
- We also know that:

$$\mu_0 = 2.5 \quad \bar{x} = 1.607$$
$$s = 1.892 \quad n = 28$$

- Then

$$t_{data} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.607 - 2.4}{1.892/\sqrt{28}} = -2.2178$$

- Since $t_{data} < 0$, we have

$$\mathrm{p-value} = 2 \cdot P(t < t_{data}) = 2 \cdot P(t < -2.2178) = 2 \cdot 0.01758 = 0.035$$

- Since $\mathrm{p-value} = 0.035 < \alpha = 0.05$ we reject H_0
- At $\alpha = 0.05$ level of significance, there is evidence that the population mean number of Customer Service Calls of such customers differs from 2.4

Assessing the strength of evidence against the null hypothesis

- If we have chosen $\alpha = 0.01$ in the prior example we would have not rejected the H_0 (since $\mathrm{p-value} = 0.035 > \alpha = 0.01$)
- Hypothesis testing restricts us to a simple "yes-or-no" decision: reject H_0 or do not reject H_0
  - It provides no indication of the strength of evidence against the null hypothesis
- Examples
  - If for a level of significance of $\alpha = 0.01$, once dataset yields p-value=0.06 and another dataset results in p-value=0.096, both p-values would lead to the same conclusion – we do not reject H_0
  - Still one dataset shows a fair amount of evidence against the null hypothesis
  - The other dataset shows no evidence at all against the null hypothesis
  - Simple "yes-or-no" decision misses the distinction between these two scenarios
  - The p-value provides extra information that the hypothesis test conclusion fails to offer

Assessing the strength of evidence against the null hypothesis *(cont'd)*

- Some data analysts think in terms of *assessing the strength of evidence against the null hypothesis*
- Thus, for the example with H_0: $\mu = 2.4$ $vs.$ H_a: $\mu \neq 2.4$, and $\mathrm{p-value} = 0.035$
  - We would not provide a conclusion
  - We would state that there is *solid evidence against the null hypothesis*

p-Value	Strength of evidence against H_0
p-value ≤ 0.001	Extremely strong evidence
$0.001 <$ p-value ≤ 0.01	Very strong evidence
$0.01 <$ p-value ≤ 0.05	Solid evidence
$0.05 <$ p-value ≤ 0.10	Mild evidence
$0.10 <$ p-value ≤ 0.15	Slight evidence
$0.15 <$ p-value	No evidence

Using confidence intervals to perform hypothesis tests

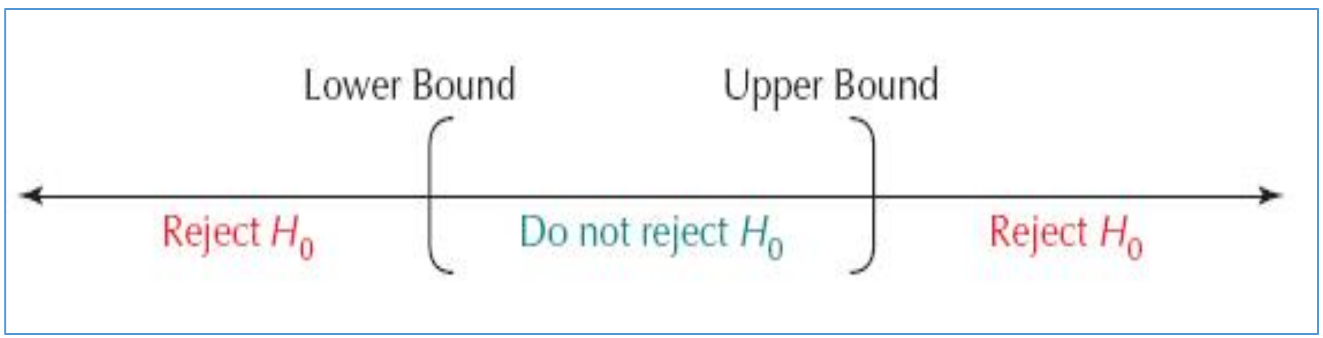
- One confidence interval is worth 1000 hypothesis tests
- The t confidence interval and the t hypothesis test are both based on the same distribution with the same assumptions

A $100(1-\alpha)\%$ confidence interval for μ is equivalent to a two-tailed hypothesis test for μ, with level of significance α.

Confidence level $100(1-\alpha)\%$	Level of significance α
90%	0.10
95%	0.05
99%	0.01

Using confidence intervals to perform hypothesis tests (*cont'd*)

- The equivalency in Figure 4.3 states that for a confidence level of $100(1-\alpha)$, if μ_0 falls:
  - **Outside of the Confidence Interval** – then the two tailed hypothesis test will reject H_0 for that value of μ_0
  - **Inside of the Confidence Interval** – then the two tailed hypothesis test will not reject H0 for that value of μ_0



Using confidence intervals to perform hypothesis tests (*cont'd*)

- Example: The 95% confidence interval for the population mean of customaer service calls was:

  (lower bound, upper bound) = (0.875, 2.339)

- Perform two-tailed confidence interval test for the following values of μ_0

  A. $\mu_0=0.5$ B. $\mu_0=1.0$ C. $\mu_0=2.4$

- The solution requires the following hypothesis tests

  A. $H_0\colon \mu = 0.5 \;\; vs. \;\; H_a\colon \mu \neq 0.5$

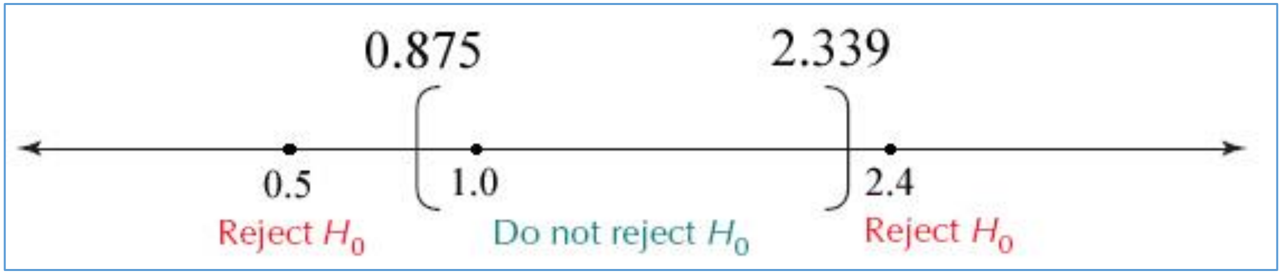
  B. $H_0\colon \mu = 1.0 \;\; vs. \;\; H_a\colon \mu \neq 1.0$

  C. $H_0\colon \mu = 2.4 \;\; vs. \;\; H_a\colon \mu \neq 2.4$

- Place the proposed values of μ_0 in the number line to determine their position with respect to the the 95% confidence interval of this variable (see Figure 4.4 in next slide)

Using confidence intervals to perform hypothesis tests (*cont'd*)

- Placement confirms that only μ_0=1.0 is within the 95% confidence interval for the two-tailed hypothesis test with level of significance α=0.05



μ_0	Hypotheses with $\alpha = 0.05$	Position in relation to 95% confidence interval	Conclusion
0.5	H_0: $\mu = 0.5$ $vs.$ H_a: $\mu \neq 0.5$	Outside	Reject H_0
1.0	H_0: $\mu = 1.0$ $vs.$ H_a: $\mu \neq 1.0$	Inside	Do not reject H_0
2.4	H_0: $\mu = 2.4$ $vs.$ H_a: $\mu \neq 2.4$	Outside	Reject H_0

Hypothesis testing for the proportion

- Hypothesis test also applies to the population proportion π as:

$$Z_{data} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

where π_0 is the hypothesized value of π and *p* is the sample proportion

$$p = \frac{\text{number of successes}}{n}$$

- Hypotheses and p-values are shown in Table 4.8

Hypotheses with $\alpha = 0.05$	p-Value
Left-tailed test. $H_0\colon \pi \geq \pi_0 \;\; vs.\; H_a\colon \pi < \pi_0$	$P(Z < Z_{data})$
Right-tailed test. $H_0\colon \pi \leq \pi_0 \;\; vs.\; H_a\colon \pi > \pi_0$	$P(Z > Z_{data})$
Two-tailed test. $H_0\colon \pi = \pi_0 \;\; vs.\; H_a\colon \pi \neq \pi_0$	If $Z_{data} < 0$ then p-value = $2 \cdot P(Z < Z_{data})$. If $Z_{data} > 0$ then p-value = $2 \cdot P(Z > Z_{data})$.

Hypothesis testing for the proportion

- Example: As 483 of 3333 customer churned, the estimate of the population proportion π of all the customers that churned is:

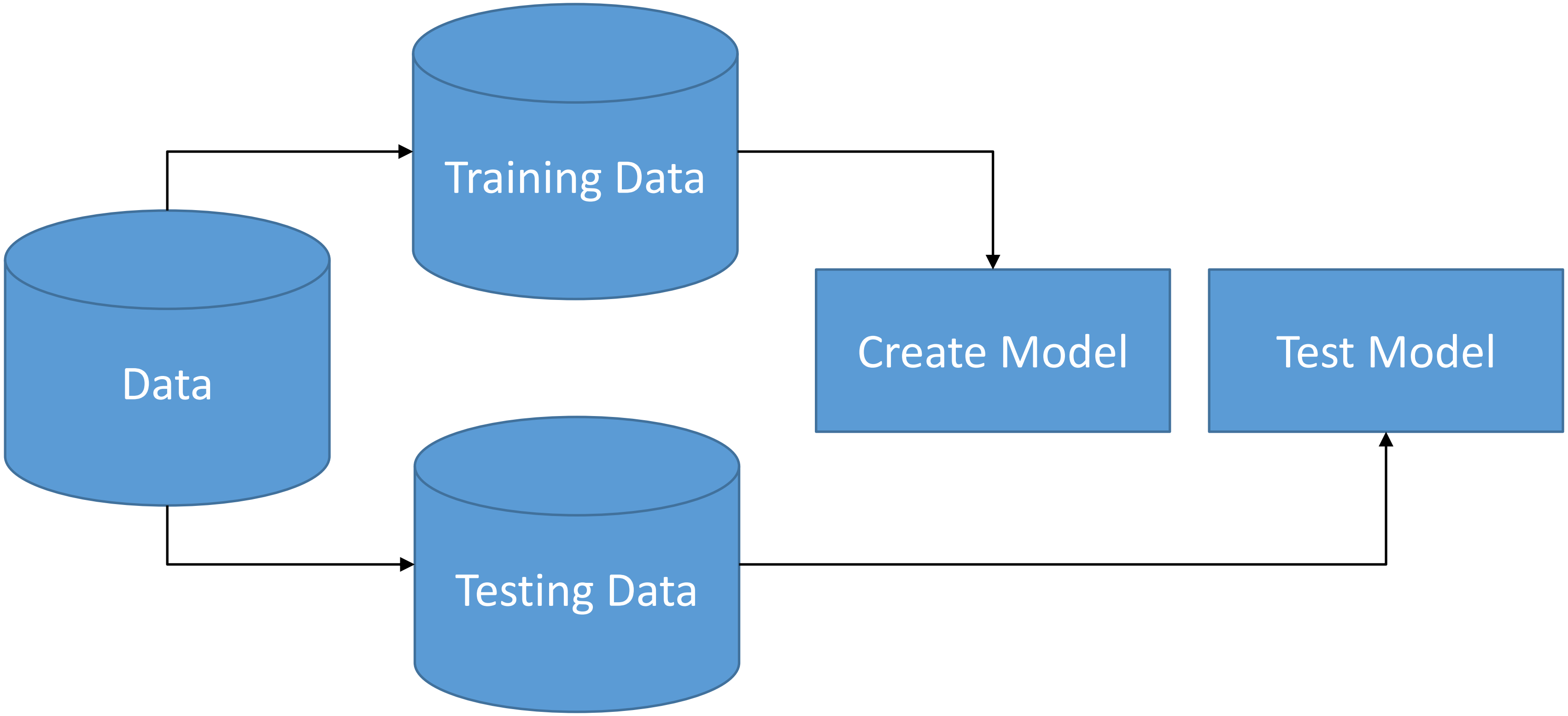
$$p = \frac{\text{number who churn}}{\text{sample size}} = \frac{x}{n} = \frac{483}{3333} = 0.1449$$

- Testing at α=0.10 whether π differs from 0.15, the hypotheses are:
$$H_0: \pi = 0.15 \;\; vs. \; H_a: \pi \neq 0.15$$
- The test statistics is:
$$Z_{data} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.1449 - 0.15}{\sqrt{\frac{0.15(0.85)}{3333}}} = -0.8246$$
- As $Z_{data} < 0$, the $\text{p} - \text{value} = 2 \cdot P(Z < Z_{data})$
  =$2 \cdot P(Z < -0.8246) = 2 * 0.2048 = 0.4096$
- Since the p-value is not less than α=0.10 we would not reject H_0

Multivariate Analysis

- Important when splitting dataset into training and test data sets
  - Bivariate hypothesis tests shown here can be used to determine whether significant differences exist between the means of various variables in the training and test data sets
  - If such differences exists, training set is not representative of the test set
  - For a continuous variable, use the two-sample t test for the difference in means.
  - For a flag variable, use the two-sample Z test for the difference in proportions.
  - For a multinomial variable, use the test for the homogeneity of proportions.
- Even if the data set has more than two variables, spot-checking of a few randomly chosen variables is usually sufficient.

Supervised Learning



Two-sample t Test for difference in means

- The test statistic for the difference in population mean is:

$$t_{data} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which follows an approximate *t* distribution with degrees of freedom the smaller of $n_1 - 1$ and $n_2 - 1$ when both populations are normally distributed or both samples are large

- Example: We divided the churn dataset into a training and a test data set
  - Assess the validity of the partition by testing whether the population mean number of customer service calls differs between the two data sets

Data Set	Sample Mean	Sample Standard Deviation	Sample Size
Training Set	$\bar{x}_1 = 1.5714$	$s_1 = 1.3126$	$n_1 = 2529$
Test Set	$\bar{x}_2 = 1.5361$	$s_2 = 1.3251$	$n_2 = 804$

Two-sample t Test for difference in means (*cont'd*)

- The sample means in Table 5.1 not too different
- Need to perform hypothesis test to make sure
- Hypothesis is:

$$H_0: \mu_1 = \mu_2 \ \ vs. \ H_a: \mu_1 \neq \mu_2$$

- The test statistic is:

$$t_{data} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1.5714 - 1.5361}{\sqrt{\frac{1.3126^2}{2529} + \frac{1.3251^2}{804}}} = 0.6595$$

- The two-tailed p-value for $t_{data} = 0.6594$ is:

$$\text{p–value} = 2 \cdot P(t > 0.6595) = 0.5098$$

- p-value is large
- There is no evidence that mean number of customer service calls differs between test and training data sets
- For this variable, the partition seems valid

Two-sample Z Test for difference in proportions

- Not all variables are numeric
- For a flag variable (like 1/0) we need the two-sample Z test for the difference in proportions

$$Z_{data} = \frac{p_1 - p_2}{\sqrt{p_{pooled} \cdot (1 - p_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $p_{pooled} = \frac{x_1+x_2}{n_1+n_2}$, and x_i and p_i represents the number of and proportion of records with value 1 (for example) for sample *i*, respectively

- Example: The Training partition in the previous example resulted in x_1=707 of n_1=2529 customers belonging to Voice Mail Plan, while the Test set has x_2=215 of n_2=804
- Therefore, $p_1 = \frac{x_1}{n_1} = \frac{707}{2529} = 0.2796$, $p_2 = \frac{x_2}{n_2} = \frac{215}{804} = 0.2674$ and $p_{pooled} = \frac{x_1+x_2}{n_1+n_2} = \frac{707+215}{2529+804} = 0.2766$
- The hypotheses are:

$$H_0\colon \pi_1 = \pi_2 \;\; vs.\; H_a\colon \pi_1 \neq \pi_2$$

Two-sample Z Test for difference in proportions (*cont'd*)

- The test statistic is:

$$Z_{data} = \frac{p_1 - p_2}{\sqrt{p_{pooled} \cdot (1 - p_{pooled}) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.2796 - 0.2674}{\sqrt{0.2766 \cdot (0.7234) \left(\frac{1}{2529} + \frac{1}{804}\right)}} = 0.6736$$

- The p-value is:

$$\text{p–value} = 2 \cdot P(Z > 0.6736) = 0.5006$$

- There is no evidence that the proportion of Voice Mail Plan differs between the training and test data sets.
- For this variable, the partition is valid

Test for the homogeneity of proportions

- *Multinomial data* is an extension of binomial data to k > 2 categories
  - Example multinomial variable: *marital status* can be *married*, *single*, *other*
  - Training set of 1000 people and test set of 250 people
- Test for the homogeneity of proportions
  - To determine whether significant differences exist between multimodal proportions
- Hypotheses are:

$$H_0\colon p_{married,training} = p_{married,test},$$
$$p_{single,training} = p_{single,test},$$
$$p_{other,training} = p_{other,test}$$

H_a: At least one of the claims in H_0 is wrong.

Data Set	Married	Single	Other	Total
Training set	410	340	250	1000
Test set	95	85	70	250
Total	505	425	320	1250

Test for the homogeneity of proportions (*cont'd*)

- Compare observed frequencies against expected frequencies if H_0 were true
- Example:
  - A. Find overall proportion of married people in whole dataset (training+test sets): ${}^{505}/_{1250}$
  - B. Multiply this overall proportion by the number of people in training set, 1000, yields the expected proportion of married people in the training set to be:

$$\text{Expected frequency}_{married,training} = \frac{(1000)(505)}{1250} = 404$$

- Step A above uses the overall proportion because H_0 states that both partitions are equal

Test for the homogeneity of proportions (*cont'd*)

- Generalizing, the expected frequencies are calculated as follows:

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

- Observed frequencies (O) and expected frequencies (E) are compared using test statistics from the χ^2_{data} (chi-square distribution:

$$\chi^2_{data} = \sum \frac{(O - E)^2}{E}$$

Data Set	Married	Single	Other	Total
Training set	404	340	256	1000
Test set	101	85	64	250
Total	505	425	320	1250

Test for the homogeneity of proportions (*cont'd*)

- Large differences between observed and expected frequencies, and large value for χ^2_{data}, leads to small p-value, and rejection of null hypothesis
- The p-value is the area to the right of χ^2_{data} under the χ^2 curve with degrees of freedom equal to (number of rows – 1)(number of columns – 1) = (1)(2) = 2:

$$p - value = P\left(\chi^2 > \chi^2_{data}\right) = P(\chi^2 > 1.15) = 0.5627$$

- Because this p-value is large, there is no evidence that the frequencies significantly differ between the training and the test data sets
- The partition is valid

Test for the homogeneity of proportions (*cont'd*)

Table 5.4

Cell		Observed Frequency	Expected Frequency	$\frac{(Obs-Exp)^2}{Exp}$
Married	Training	410	404	$\frac{(410-404)^2}{404}=0.09$
Married	Test	95	101	$\frac{(95-101)^2}{101}=0.36$
Single	Training	340	340	$\frac{(340-340)^2}{340}=0$
Single	Test	85	85	$\frac{(85-85)^2}{85}=0$
Other	Training	250	256	$\frac{(250-256)^2}{256}=0.14$
Other	Test	70	64	$\frac{(70-64)^2}{64}=0.56$
				χ^2_{data}=1.15

Chi-square test for goodness of fit of multinomial data

- Let's assume that marital status of the population is married=40%, single=35%, other=25%
- Determine whether the sample is representative of the population
- Use χ^2 (chi-square) goodness of fit test. Hypotheses are:

$$H_0\colon p_{married} = 0.40, \qquad p_{single} = 0.35, \qquad p_{married} = 0.25$$

H_a: At least one of the proportions in H_0 is wrong

- Sample size n=100 yields the following observed frequencies

$$O_{married} = 36, \qquad O_{single} = 35, \qquad O_{other} = 29$$

- Need to compare observed frequencies against the expected frequencies assuming that H_0 is true

$$E_{married} = n \cdot p_{married} = 100 \cdot 0.40 = 40$$
$$E_{single} = n \cdot p_{single} = 100 \cdot 0.35 = 35$$
$$E_{other} = n \cdot p_{other} = 100 \cdot 0.25 = 25$$

Chi-square test for goodness of fit of multinomial data *(cont'd)*

- Comparing the frequencies using the test statistic

$$\chi^2_{data} = \sum \frac{(O-E)^2}{E}$$

- Large differences between observed and expected frequencies, a large value for χ^2_{data} will lead to a small p-value, and rejection of null hypothesis.

Cell	Observed Frequency	Expected Frequency	$\frac{(Obs-Exp)^2}{Exp}$
Married	36	40	$\frac{(36-40)^2}{40} = 0.4$
Single	35	35	$\frac{(35-35)^2}{35} = 0$
Other	29	25	$\frac{(29-25)^2}{25} = 0.64$
			χ^2_{data}=1.04

Chi-square test for goodness of fit of multinomial data (*cont'd*)

- The p-value is the area to the right of χ^2_{data} under the χ^2 curve with k-1 degrees of freedom, where k-number of categories (here k=3)

$$p - value = P\left(\chi^2 > \chi^2_{data}\right) = P(\chi^2 > 1.04) = 0.5945$$

- There is no evidence that the observed frequencies represent proportions that differ from those in the null hypothesis
- Our sample is representative of the population

Analysis of Variance

- One-way analysis of variance (ANOVA)
- Used as an extension of the situation for the two-sample t test
- For example: Check whether mean value of a continuous variable is the same across three partitions of a data set
- For the samples for Age variable from Groups A, B and C, we set the following hypotheses

$$H_0\colon\ \mu_A = \mu_B = \mu_c$$

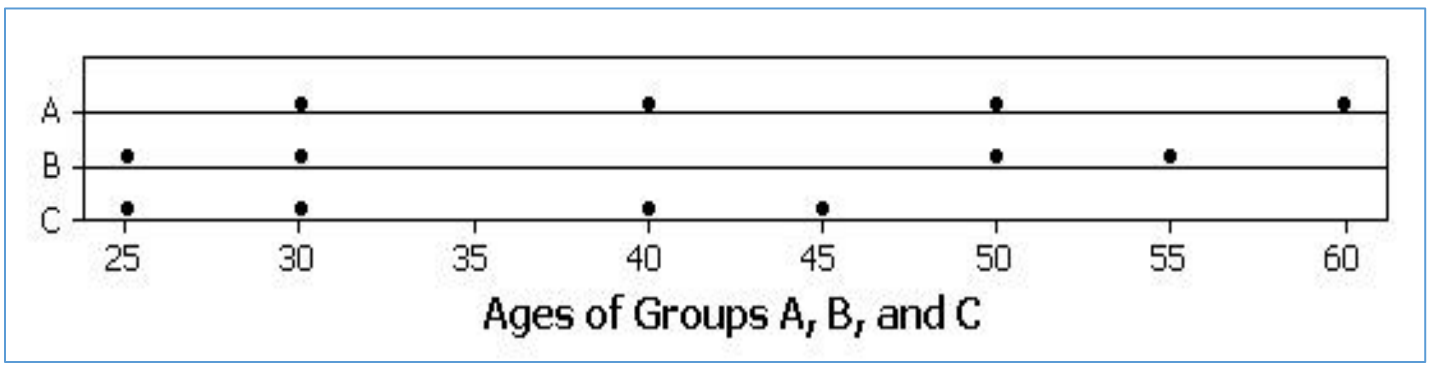
$$H_a\colon \text{Not all the population means are equal}$$

- Sample mean ages are $\bar{x}_A = 45$, $\bar{x}_B = 40$, and $\bar{x}_C = 35$

Group A	Group B	Group C
30	25	25
40	30	30
50	50	40
60	55	45

Analysis of Variance (*cont'd*)

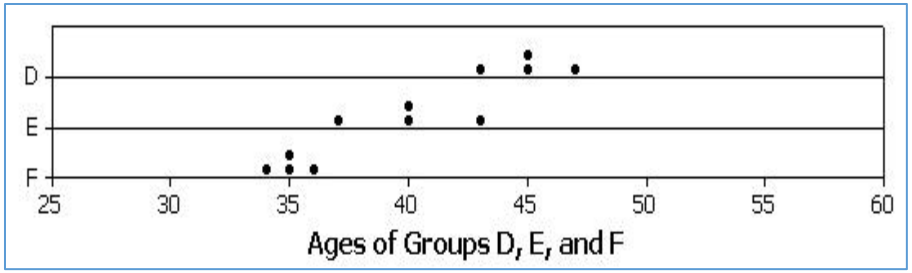
- Dot plot of the data shows considerable amount of overlap among the groups
- Dot plot offers little evidence to reject the null hypotheses that the population means are all equal



Analysis of Variance (*cont'd*)

- Consider now the groups D, E and F
- Sample means are again $\bar{x}_D = 45$, $\bar{x}_E = 40$, and $\bar{x}_F = 35$
- But the dot plot shows little overlap among data sets
  - There is good evidence to reject the null hypotheses

Group D	Group E	Group F
43	37	34
45	40	35
45	40	35
47	43	36



Analysis of Variance (*cont'd*)

- When spread within each sample is large, the difference in sample means seems small
- When the spread within each sample is small, the difference in sample means seems large
- Analysis of variance work by comparing:
    1. Between-sample variability (measured by the variability of the sample means)
    2. Within-sample variability (measured, for example, by the sample standard deviation)
- When #1 is much larger than #2, this represent evidence that the population means are not equal.
- The analysis depends on measuring variability, hence the name *analysis of variance*

Analysis of Variance (*cont'd*)

- Let $\bar{\bar{x}}$ represent the overall sample mean
- Measure the between-sample variability by finding the variance of the *k* sample means, weighted by sample size
  - Expressed as the *mean square treatment* (MSTR):

$$MSTR = \frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

- Measure the within-sample variability by finding the weighted mean of the sample variances
  - Expressed as the *mean square error* (MSE):

$$MSE = \frac{\sum (n_i - 1) s_i^2}{n_t - k}$$

- We compare these two quantities by taking their ratio:

$$F_{data} = \frac{MSTR}{MSE}$$

which follows an *F* distribution, with degrees of freedom $df_1 = k - 1$ and $df_2 = n_t - k$

Analysis of Variance (*cont'd*)

- The numerator of MSTR is the *sum of squares treatment*, SSTR
- The numerator of MSE is the *sum of squares error*, SSE.
- The total sum of squares (SST) is the sum of SSTR and SSE
- ANOVA table is a convenient way of displaying this information
- F_{data} will be large when the between-sample variability is much greater than the within-sample variability – this calls for rejection of null hypothesis
- The p-value is $P(F > F_{data})$; reject the null hypothesis when the p-value is small

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Treatment	SSTR	$df_1 = k - 1$	$MSTR = \frac{SSTR}{df_1}$	$F_{data} = \frac{MSTR}{MSE}$
Error	SSE	$df_2 = n_t - k$	$MSE = \frac{MSE}{df2}$	
Total	SST			

Analysis of Variance (*cont'd*)

- Example:ANOVA results for Groups A, B, and C/Groups D, E, and F
- The p-value=0.548 indicates that there is no evidence against the null hypothesis that all population means are equal
- The p-value=0.000 indicates that there is strong evidence that not all population means ages are equal, supporting the claim

Group A	Group B	Group C
30	25	25
40	30	30
50	50	40
60	55	45

Group D	Group E	Group F
43	37	34
45	40	35
45	40	35
47	43	36

Source	DF	SS	MS	F	P
Factor	2	200	100	0.64	0.548
Error	9	1400	156		
Total	11	1600			

Source	DF	SS	MS	F	P
Factor	2	200.00	100.00	32.14	0.000
Error	9	28.00	3.11		
Total	11	228.00			

Regression Analysis

- In this section we use the *Cereals** data set
- Data set is included in the book series web site

Cereals dataset variables	
Cereal manufacturer	Grams of sugars
Type (hot or cold)	Milligrams of potassium
Calories per serving	Percentage of recommended daily allowance of vitamins (0% 25%, or 100%)
Grams of protein	Weight of one serving
Grams of fat	Number of cups per serving
Milligrams of sodium	Shelf location (1 = bottom, 2 = middle, 3 = top)
Grams of fiber	Nutritional rating, calculated by Consumer Reports
Grams of carbohydrates	

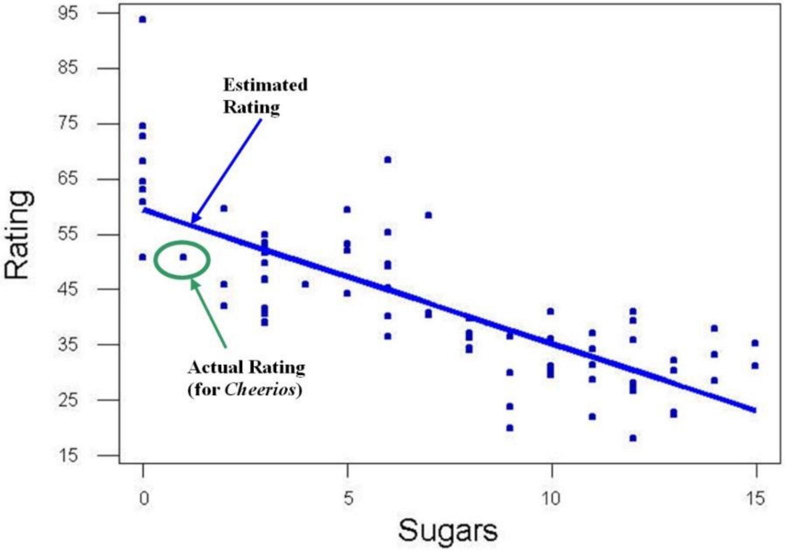
Regression Analysis (*cont'd*)

- Caveat: This dataset contains missing data:
  - Potassium content of Almond Delight
  - Potassium content of Cream of Wheat
  - Carbohydrates and sugars content of Quaker Oatmeal
- Will not be able to use sugar content in Quaker Oatmeal to build model

Cereal Name	Manuf.	Sugars	Calories	Protein	Fat	Sodium	Rating
100% Bran	N	6	70	4	1	130	68.4030
100% Natural Bran	Q	8	120	3	5	15	33.9837
All-Bran	K	5	70	4	1	260	59.4255
All-Bran Extra Fiber	K	0	50	4	0	140	93.7049
Almond Delight	R	8	110	2	2	200	34.3848
Apple Cinnamon Cheerios	G	10	110	2	2	180	29.5095
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Regression Analysis (*cont'd*)

- Scatter plot of the nutritional rating vs sugar content, along with the least-squares regression line
- The regression equation is $\hat{y} = b_0 + b_1 x$, where:
  - $\hat{y}$ is the estimated value of the response variable
  - b_0 is the y-intercept of the regression line
  - b_1 is the slope of the regression line
  - b_0 and b_1, together, are called the regression coefficients



Regression Analysis (*cont'd*)

- The top regression equation Rating=59.9-2.46 Sugars shows rounded coefficients. More digits in the "Coef" column
  - b_0=59.852
  - b_1=-2.4614
- Thus, $\hat{y} = 59.852 - 2.4614x$
- Interpreted as "*The estimated cereal rating equals 59.853 minus 2.4614 times the sugar content in grams*"
- The regression line and equations are linear approximations of the relationship between x (predictor) and y (response) variables – sugar and nutritional rating in this case
- Regression equation can be used to make estimates and predictions

```
The regression equation is
Rating = 59.9 - 2.46 Sugars


76 cases used, 1 cases contain missing values


Predictor      Coef   SE Coef       T      P
Constant     59.853     1.998   29.96  0.000
Sugars      -2.4614    0.2417  -10.18  0.000


S = 9.16616   R-Sq = 58.4%   R-Sq(adj) = 57.8%


Analysis of Variance

Source          DF       SS      MS       F      P
Regression       1   8711.9  8711.9  103.69  0.000
Residual Error  74   6217.4    84.0
Total           75  14929.3


Unusual Observations

Obs  Sugars  Rating    Fit  SE Fit  Residual  St Resid
  1     6.0   68.40  45.08    1.08     23.32     2.56R
  4     0.0   93.70  59.85    2.00     33.85     3.78R

R denotes an observation with a large standardized residual.


Predicted Values for New Observations

New Obs    Fit  SE Fit       95% CI            95% PI
      1  57.39    1.80  (53.81, 60.97)  (38.78, 76.00)


Values of Predictors for New Observations

New Obs  Sugars
      1    1.00
```

Regression Analysis (*cont'd*)

- Example: Estimate the nutritional rating for <u>new cereal</u> that contains x=1 gram of sugar
  - Using the regression equation: $\hat{y} = 59.852 - 2.4614(1) = 57.3916$
  - Estimated value lies directly on the regression line, at (x=1, $\hat{y} = 57.3916$)
  - For any value of x (sugar content) the estimated value for y (nutritional rating) lies precisely in the regression line
- Cheerios cereal also has sugar=1 gram, but rating is 50.765, not the estimated 57.3916. Point (x=1, $\hat{y} = 50.765$)
- Prediction using regression line deviated by 50.765 – 57.3916 = –6.6266 rating points
  - This vertical distance in general ($y - \hat{y}$) is known as *prediction error, estimation error*, or *residual*
- Two unusual observations: cereal 1 (100% Bran) and cereal 4 (All-Bran with Extra Fiber), which have large positive residuals, indicating nutrition rating unexpectedly high, given their sugar level

Regression Analysis (*cont'd*)

- Regression objective is to minimize the overall size of the prediction errors
  - Least-squares regression works by choosing the regression that minimizes the sum of squared errors (SSE) – this is the most common method
  - Median regression, is an alternative method
- y-intercept b_0 is the location on the y-axis where the regression line intercepts the y-axis; the value when the predictor is zero
  - However, in many regressions a predictor with value of zero would not make sense
  - Example: A regression predicting the weight of an elementary school student based on height would not make sense when height=0
  - However, for the cereal data set, b_0 represents the nutritional content of a cereal with no sugar
- The slope of the regression line indicates the change in y per unit increase in x
  - $b_1 = -2.4614$ means, "For each increase of 1 gram in sugar content, the estimated nutritional rating decreases by 2.4614 rating points"
  - For example: For cereal A with 5 more grams of sugar than cereal B would have estimated nutritional rating of 5(2.4614)=12.307 rating points lower than cereal B.

Hypothesis testing in regression

- Objective is to use the known value of the slope b_1 to perform inference for the unknown value of the slope β_1 of the population regression equation
  - Just like when in univariate analysis we used sample mean $\bar{x}$ to infer the unknown value of the population mean μ
- The population regression equation represents the relationship between, say, nutritional ratings and sugar content for the entire population of cereals – not just the ones in the sample

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε represents a random variable for modeling the errors

- Notice that when β_1=0, the population equations becomes $y = \beta_0 + \varepsilon$ and there is no relationship between the predictor x and the response y
- For any other value of β_1 there is a linear relationship between x and y

Hypothesis testing in regression (*cont'd*)

- We wish to test for the existence of a linear relationship between *x* and *y*
- Thus, perform the following hypothesis test

$H_0: \beta_1 = 0$ No relationship between *x* and *y*

$H_a: \beta_1 \neq 0$ Linear relationship between *x* and *y*

- The test statistic for this hypothesis test is:

$$t = \frac{b_1}{s_{b_1}}$$

where s_{b_1} represents the standard error for b_1

- Large values of s_{b_1} indicated too much variability in the slope of b_1 – this makes precise inference difficult
- Per the formula above, large values of s_{b_1} reduce the size of the *t-statistic*

Measuring the quality of a regression model

- If we cannot reject the null hypothesis that $\beta_1 = 0$, the regression is not useful
- If we find that $\beta_1 \neq 0$, then there are two statistics to measure the quality of the regression
  - Standard error of the estimate: *s*
  - R-squared: r^2
- Standard error of the estimate: s
  - Should not be confused with *s*, the sample standard deviation for univariate statistics, or $s_{b_1}$the standard error of the slope coefficient
  - Defined as

$$s = \sqrt{MSE} = \sqrt{SSE/n-2}$$

  where SSE is the sum of squared errors
  - Its value indicate the size of the "typical" prediction error

Measuring the quality of a regression model (*cont'd*)

- R-squared statistic: r^2
  - Measures how closely the linear regression fits the data
  - Values near 100% indicate a more perfect fit
  - Defined as:

$$r^2 = \frac{SSR}{SST}$$

  Where SST represents the variability in the y-variable, and SSR represents the improvement in the estimation as compared to just using $\bar{y}$

  - May be interpreted as the ratio of total variability in y that is accounted for by the linear relationship between x and y

Dangers of extrapolation

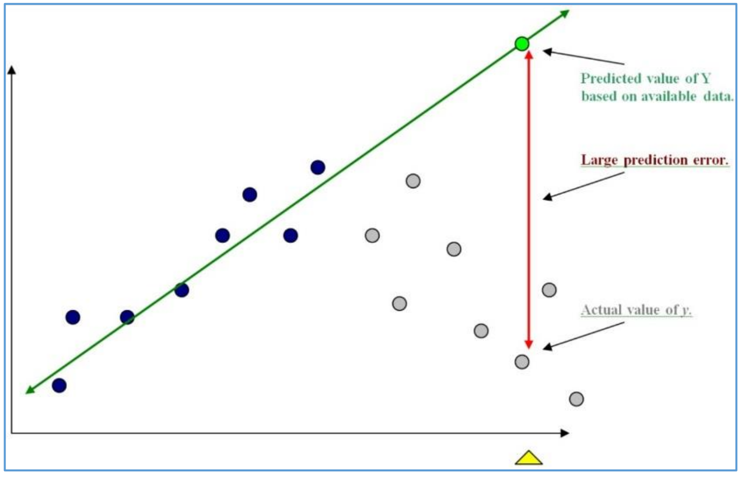
- Using the estimated regression on a new cereal with 30 grams of sugar per service yields:

$$\hat{y} = 59.853 - 2.4614(sugars) = 59.853 \ - 2.4614(30) = \mathbf{-13.989}$$

- The minimum rating in the dataset is 18, but this new cereal has a negative rating. What is going here?
- This is an example of the dangers of extrapolation
- Analyst should restrict estimates and predictions to the values within the range of the values of x in dataset
  - Example: The range of sugar content in the dataset is from 0 to 15
  - Predictions of nutritional rating for cereals with 0-15 grams of sugar be appropriate
  - Prediction outside of this range would be dangerous, since we don't know the nature of the relationship outside of this range
- Extrapolation should be avoided, and end user should be informed that no x-data is available to support a prediction
- Relationship outside of this range may no longer be linear

Dangers of extrapolation (*cont'd*)

- The green line regression was developed using available data represented by black dots
- Unobserved data at higher values of x is represented by gray dots
- If we make a prediction at the value of x where the yellow triangle is located, we would incur in a massive error as large as the red arrow
- The analyst would be unaware of the magnitude of the predicted error
- Policy recommendations based on such wrong predictions might have costly errors



Confidence intervals for the mean value of y given x

- Point estimates suffer from the lack of probability statement associated with their accuracy
- We then use confidence intervals for the man value of y for a given value of x, with the following formula

$$point\ estimate \pm margin\ of\ error = \hat{y}_p \pm t_{\propto/2}(s)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

where

x_p = the particular value of x for which the prediction is being made

$\hat{y}_p$ = the point estimate of y for a particular value of x

$t_{\propto/2}$ = a multipler associated with the sample size and confidence level

s = the standard error of the estimate

SSE = the sum of squared residuals

Predictions for a randomly chosen value of y given x

- It is 'easier' to predict a mean value than a randomly chosen value of a variable
  - Example: It is easier to predict the batting average, or the school class exam score average than predicting an individual player battling average or a specific student exam score
- Usually data miners are more interested in predicting individual value, rather than mean of all the values, given x
- Prediction intervals are used to estimate the value of a randomly chosen value of y, given x; as follows:

$$point\ estimate \pm margin\ of\ error = \hat{y}_p \pm t_{\propto/2}(s)\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x}^2}{\sum(x_i-\bar{x})^2}}$$

- This formula is the same as the formula for the confidence interval of the mean value of y, given x, excel for the presence of the "1+" inside the square root
- The "1+" ensures a wider prediction interval than the analogous confidence interval

Multiple Regression

- Some data sets include hundreds of variables, which may have a linear relationship with the target variable
- Multiple regression modeling provides a method for describing those relationships
- Example: For the *Cereal* dataset, add *sodium* (in addition to existing variable *sugar*) for predicting the *rating*, and observe whether the quality of the model improved or not
- The equation for multiple regression with two predictors is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

- Figure 5.8 shows the regression results for nutritional rating based on sugars and sodium

```
The regression equation is
Rating = 69.2 - 2.39 Sugars - 0.0606 Sodium


76 cases used, 1 cases contain missing values


Predictor       Coef   SE Coef       T      P
Constant      69.180     2.373   29.15  0.000
Sugars       -2.3944    0.2041  -11.73  0.000
Sodium      -0.06057   0.01086   -5.58  0.000


S = 7.72769   R-Sq = 70.8%   R-Sq(adj) = 70.0%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  10569.9  5285.0  88.50  0.000
Residual Error  73   4359.4    59.7
Total           75  14929.3


Source  DF  Seq SS
Sugars   1  8711.9
Sodium   1  1858.0


Unusual Observations

Obs  Sugars  Rating     Fit  SE Fit  Residual  St Resid
  1     6.0  68.403  46.940   0.970    21.463     2.80R
  2     8.0  33.984  49.116   1.845   -15.133    -2.02R
  3     5.0  59.426  41.461   1.465    17.964     2.37R
  4     0.0  93.705  60.701   1.691    33.004     4.38R

R denotes an observation with a large standardized residual.


Predicted Values for New Observations

New Obs     Fit  SE Fit         95% CI              95% PI
      1  49.222   2.107  (45.023, 53.421)  (33.258, 65.185)


Values of Predictors for New Observations

New Obs  Sugars  Sodium
      1    1.00     290
```

Multiple Regression (*cont'd*)

- Multiple regression equation is:

$$\hat{y} = 69.180 - 2.3944\ (sugars) - 0.06057\ (sodium)$$

where b_2=-0.06057 is interpreted so that for each addition milligram of sodium, the estimated decrease in nutritional rating is 0.06057, ***when sugars is held constant***

- The point estimate of rating for Cheerios, with 1 gram is sugar and 290 mg of sodium is:

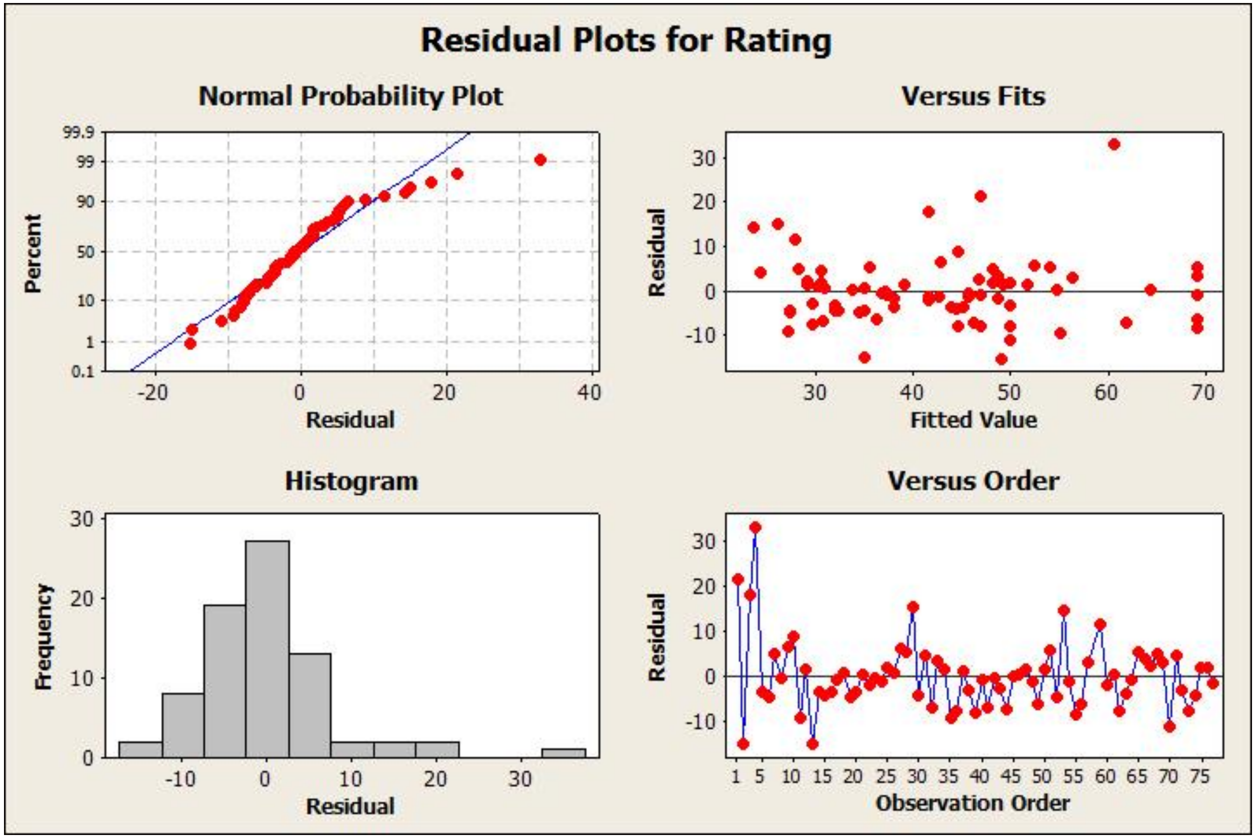
$$\hat{y} = 69.180 - 2.3944\ (1) - 0.06057\ (290) = 49.22$$

- The prediction error for Cheerios is the difference between actual rating *y* and the predicted rating $\hat{y}$: (y-$\hat{y}$)=50.765-49.22=1.545
  - This prediction error is smaller than the obtained using sugars only: $(y - \hat{y}) = -6.62266$
  - This, because our model uses double the data (two predictors instead of one)
- The standard error of the estimate has been reduced from s≈9.2 to s≈7.7
- The value for r^2 has increased from 58.4% to 70.8%
  - 70.8% of our variability in nutritional rating is explained by our regression model

Verifying model assumptions

- Before implementing a model, the requisite model assumptions must be verified
- Making predictions using a model were assumptions are violated may lead to erroneous results
- The assumptions are: *Linearity*, *Independence*, *Normality* and *Constant Variance*
- Assumptions might be checked with:
  - Normality plot of residuals (Figure 5.9, upper left)
    - Checks whether there are systematic deviations from linearity
    - In there are deviations, the values do not follow a normal distribution
  - Plot of standardized residuals against the fitted (predicted) values
    - If obvious curvature exists in the scatter plot, the linearity assumption is violated
    - If the vertical spread is systematically non-uniform, the constant variance assumption is violated

Verifying model assumptions (*cont'd*)



Verifying model assumptions (*cont'd*)

- Important
  - We must be careful not to find patterns where no such pattern exists!
  - Departure from linearity must be systematic and significant
  - The huge data sets used in data mining do not always follow perfect normality – so that analyst should give the data the benefit of the dount unless good evidence exists for the contrary
- Outliers have an outsized influence in Least squares regression
  - Should identify **extreme** outliers, and *if necessary and appropriate*, omit them