

## Administrative Issues

Web site – [www-stat.wharton.upenn.edu/~stat102](http://www-stat.wharton.upenn.edu/~stat102)

- TEXT: Dielman, T. *Applied Regression Analysis Fourth Edition*, Duxbury
- Lecture notes will be posted to the web site – usually by the night before the lecture. You are responsible for printing them out.
- **Syllabus** is posted on the website
- Data, homework assignments and solutions will be posted on the web site.
- Office hours – L. Brown; 445 Huntsman; Tue & Thurs 10 – 11  
L. Zhao; 470 Huntsman; Mon 3-5
- Stat Lab – Huntsman F-96, tentative hours 10-5 M, T, Th, F & 10-3 W

1-1

- Computer package: JMP (any version #5.0 or higher will suffice)
- Homework and Projects (See syllabus and website for more details)
- **Exams:**
  - 6-7:30, Mon, Feb 19, & Thu, March 29, (midterm 1 & 2)
  - 6-8 May 1 (Final)
- Grading:
  - Homework and Projects – 25%,
  - Midterms – 20% Each,
  - Final – 35%.

(Lower midterm score will be replaced by final score if the latter is higher.)

- Teaching Assistants (offices in Huntsman)
  - Oliver Entine – 434-2; Hui Nie – 434-2; Osbert Pang – 431-1
  - Office Hours – TBA

### First HW Assignment

Read: Chapter 2:

Sections 2.1 through 2.7 should be “review”. Sections 2.8 & 2.9 may be “new”

Written HW: (To hand in at class on Tue, Jan 16 or Wed, Jan 17.)

Problems 2.25, 2.30, 2.28, 2.35, 2.40, 2.45 (For part **c**, comment on whether a case-control design would be more appropriate. If so, explain.), 2.52, 2.56, 2.61.

[For the problems based on a data file you should use JMP. You may find the data available on the class website to be in a more convenient format than that supplied with your book.]

1-2

## What You Should Already Know (Stat 101)

### Graphical tools

Histogram, quantile plot, boxplot, comparison boxplots, and scatterplot.

### Expected value, variance and covariance

Expected value is an average, weighted by probabilities.

Variance is the expected squared deviation from mean.

Covariance measures the strength of linear association.

### Normal distribution

95% of the distribution lies in the range  $\mu \pm 2\sigma$

Normal quantile plot as a diagnostic

### Standard error

Sample-to-sample variability of a statistic

$$SE(\bar{x}) = s/\sqrt{n}$$

### Confidence interval (for a normal mean)

Estimate  $\pm 2$  SE(Estimate) [**2** is a good approximation to exact value]

Contains “truth” for 95% of samples

1-3

### Hypothesis test (about a normal mean)

The null hypothesis,  $H_0$  is the *opposite* of what is hoped to be validated

Standardized z-statistic, or t-statistic/t-ratio, counts the SE's from conjectured value

P-value measures probability-size of t or |t| (for one or two-sided test)

The smaller the p-value, the greater the evidence against  $H_0$

P-value  $< 0.05 \iff$  reject  $H_0$  at the .05 level of significance

### Software

JMP – to the extent it was used in Stat 101

1-4

## A Quick Review

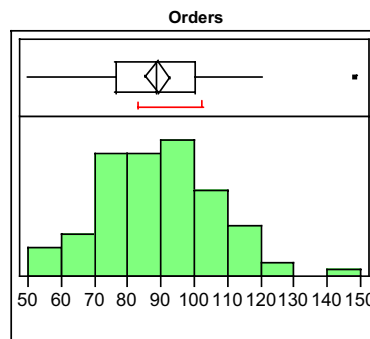


- Companies that sell groceries over the Internet are called e-grocers. Customers enter their orders, pay by credit card and receive delivery by truck.
- In order for e-grocing to be profitable, the average order size (in \$) has to be reasonably large.
- To determine the potential profitability of e-grocing, a grocer in a large city offered the service and recorded the size of the order for a random sample of 85 customers.
- The grocer wants to understand the distribution of order sizes.

PS: Typical sample sizes for such a preliminary market study would be much larger than our sample of  $n = 85$ .

1-5

## Graphical Analysis of the Data (egrocery.jmp)<sup>1,2</sup>



Boxplot: there is one outlier, an order of about \$150. Other than the outlier, the distribution is not particularly skewed.

Histogram: The distribution of order sizes is unimodal and looks roughly symmetric, other than the outlier.

<sup>1</sup> To obtain the graphical analysis and means and variances in JMP, click Analyze, Distribution and put the variable of interest in the Y-Columns box.

<sup>2</sup> The data are available on our website as egrocery.jmp

1-6

## Sample Means and Variances, and Quantiles

	Sample Moments	
Mean		$\bar{Y} = 89.27$
Std Dev		$s = 17.30$
N		85

Sample Mean:  $\bar{Y} = \frac{Y_1 + \dots + Y_n}{n}$ .

Sample Standard Deviation:  $s = \sqrt{\frac{(Y_1 - \bar{Y})^2 + \dots + (Y_n - \bar{Y})^2}{n - 1}}$ .

	Sample Quantiles	
75.0%	quartile	100.4
50.0%	median	88.7
25.0%	quartile	76.9

Inter-quartile distance:  $100.4 - 88.7 = 11.7$

1-7

## Normal Distribution (Population distribution)

If a random variable Y is normally distributed, then

About 68% of the time, a randomly drawn Y will lie within one standard deviation of its mean.

About 95% of the time, a randomly drawn Y will lie within two standard deviations of its mean.

About 99% of the time, a randomly drawn Y will lie within three standard deviation of its mean.

*Suppose* order size was normal; with a mean of 89; and the sd of 17.

Then **about** 95% of the order sizes will be between

$$89 \pm 2 * 17 = (55, 123)$$

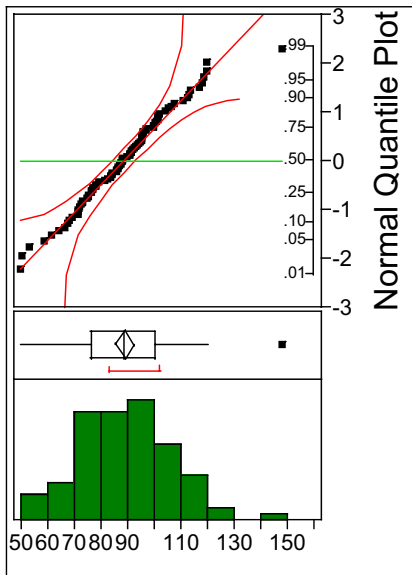
Are order sizes normally distributed?

Diagnostic tool: Normal quantile plot.<sup>3</sup> (Uses the data to see whether this is reasonable)

<sup>3</sup> To obtain the normal quantile plot, after using Analyze - Distribution to obtain the histogram, click the red triangle next to the Y variable and click Normal Quantile plot.

1-8

## Normal Quantile Plot



- If the data fit the normal model, the data points should fall approximately on the red solid diagonal line.
- Normality is a reasonable approximation for the data that excludes the outlier.

1-9

## Standard Errors and Confidence Intervals

The grocer is interested in the mean order size for the population of all potential customers, in order to determine whether e-grocery will be profitable.

Sample mean = 89.3 = *estimate* of the population mean.

Statistical Inference: Make inferences about a *population* quantity (e.g., population mean) from the sample.

Sampling distribution of an estimate: The distribution of this type of estimate over repeated samples, e.g., distribution of the sample mean over repeated samples of 85 customers.

Standard error of an estimate: Standard deviation of the sampling distribution of an estimate.

1-10

	<b>Moments</b>	
Mean		89.27
Std Dev "s"		17.30
Std Err Mean		1.88
upper 95% Mean		93.00
lower 95% Mean		85.54
N (or "n")		85

$$SE(\bar{Y}) = \frac{s}{\sqrt{n}} = \frac{17.3}{\sqrt{85}} = 1.88$$

95% Confidence Interval (**approximate**):

Estimate  $\pm 2$  SE(Estimate)

Contains "truth" for 95% of samples

95% Confidence Interval for mean order size

$$\approx 89.27 \pm 2 * 1.88 = (85.51, 93.03)$$

Exact 95% confidence interval *assuming the population distribution is normal* is given in JMP by "lower 95% mean" and "upper 95% mean".

Under these assumptions the 95% CI has prob = 0.95 of containing the true pop. mean.

PS: The above calculations *include* the outlier.

1-11

## Hypothesis Testing

The grocer determines that in order for e-grocery to be profitable, the mean order size needs to be greater than \$85.

Hypothesis testing: Decide between two hypotheses, the null hypothesis and the alternative hypothesis. The alternative hypothesis (sometimes called research hypothesis) is the hypothesis that you are hoping to establish to be true.

$$H_0 : \mu_{order\_size} \leq 85$$

$$H_a : \mu_{order\_size} > 85$$

t-statistic/t-ratio counts the SE's from conjectured value

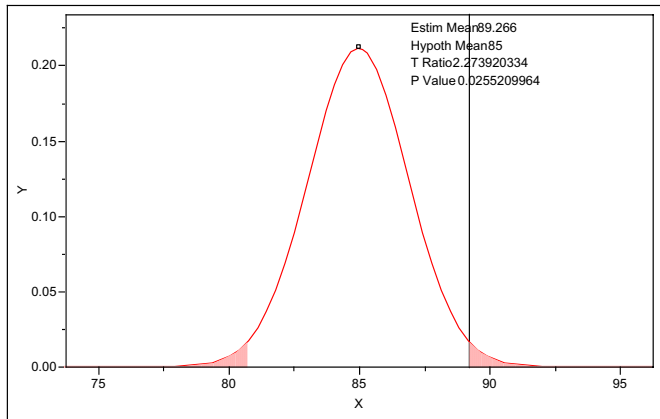
p-value  $< 0.05 \Leftrightarrow$  reject  $H_0$  at the .05 level of significance

$$t\text{-statistic} = \frac{\bar{Y} - 85}{SE(\bar{Y})} = \frac{89.3 - 85}{1.88} = 2.29$$

P-value = Probability if the null hypothesis were true that t-statistic would provide at least as much as evidence as the observed t-statistic.

1-12

### JMP Output for Hypothesis test



<b>Test Mean=value</b>	
Hypothesized Value	85
Actual Estimate	89.27
Df	84
Std Dev	17.30
	<b>t Test</b>
Test Statistic	2.274
Prob >  t	0.026
Prob > t	0.013
Prob < t	0.987

p-value = 0.013.  $\Rightarrow$  Decision is to “REJECT  $H_0$  at level  $\alpha = 0.05$ ”

p-value < 0.05 means that there is evidence (at level  $\alpha = 0.05$ ) that the alternative hypothesis is true – there is evidence that the mean order size is >\$85.

**The grocer decides that e-grocery will be profitable.**