CS 660: Mathematical Foundations for Analytics

Dr. Francis Parisi

Pace University

Spring 2017

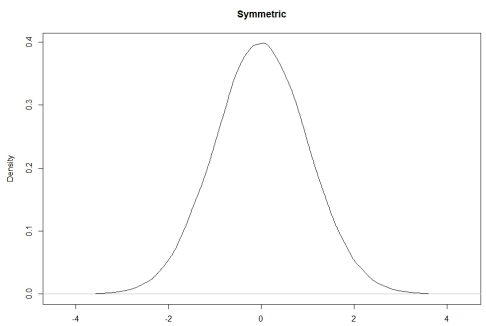
Course Overview

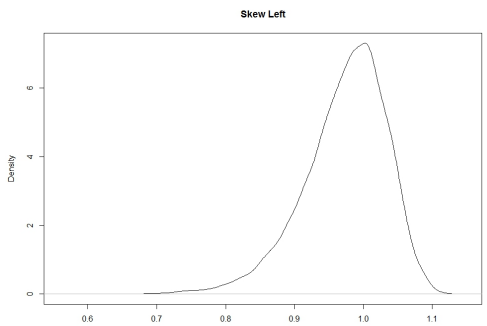
1. Part I: Data Science Fundamentals
   - Data Science Concepts and Process
   - The R Language
   - Exploratory Data Analysis
   - Cleaning & Manipulating Data
   - Presenting Results
2. Part II: Graphs & Statistical Methods
   - Basic Graphics
   - Advanced Graphics
   - Probability & Statistical Methods
3. Part III: Modeling Methods
   - Linear and Logistic Regression
   - Model Selection and Evaluation
   - Unsupervised Methods
   - Advanced Modeling Methods

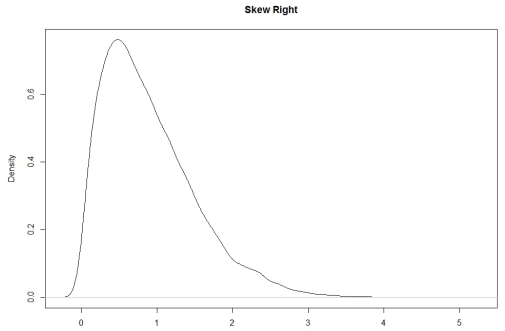
Probability & Statistical Methods

Descriptors for probability distributions...

- Location – the center or the peak of the distribution typically measured by the mean (average or expected value) or median
- Shape – the degree of symmetry or skewness



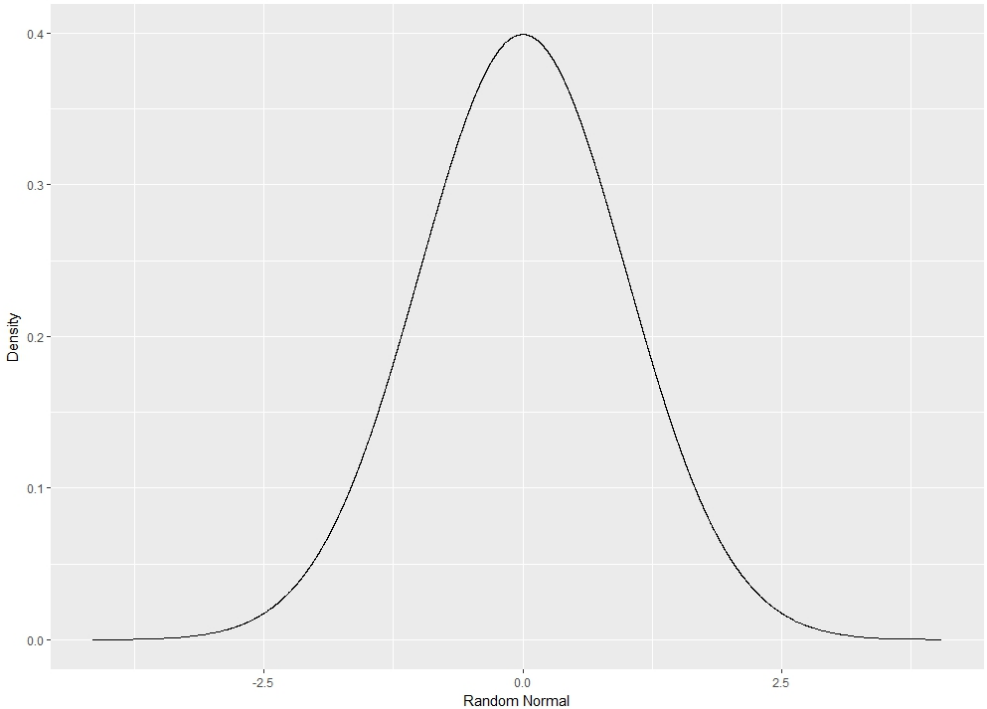




- Scale or Variability – the spread of the distribution typically measured by variance or standard deviation, the range or interquartile range, or a scale parameter
- Deviations – possible outliers, or unusual points that are not consistent with the rest of the data

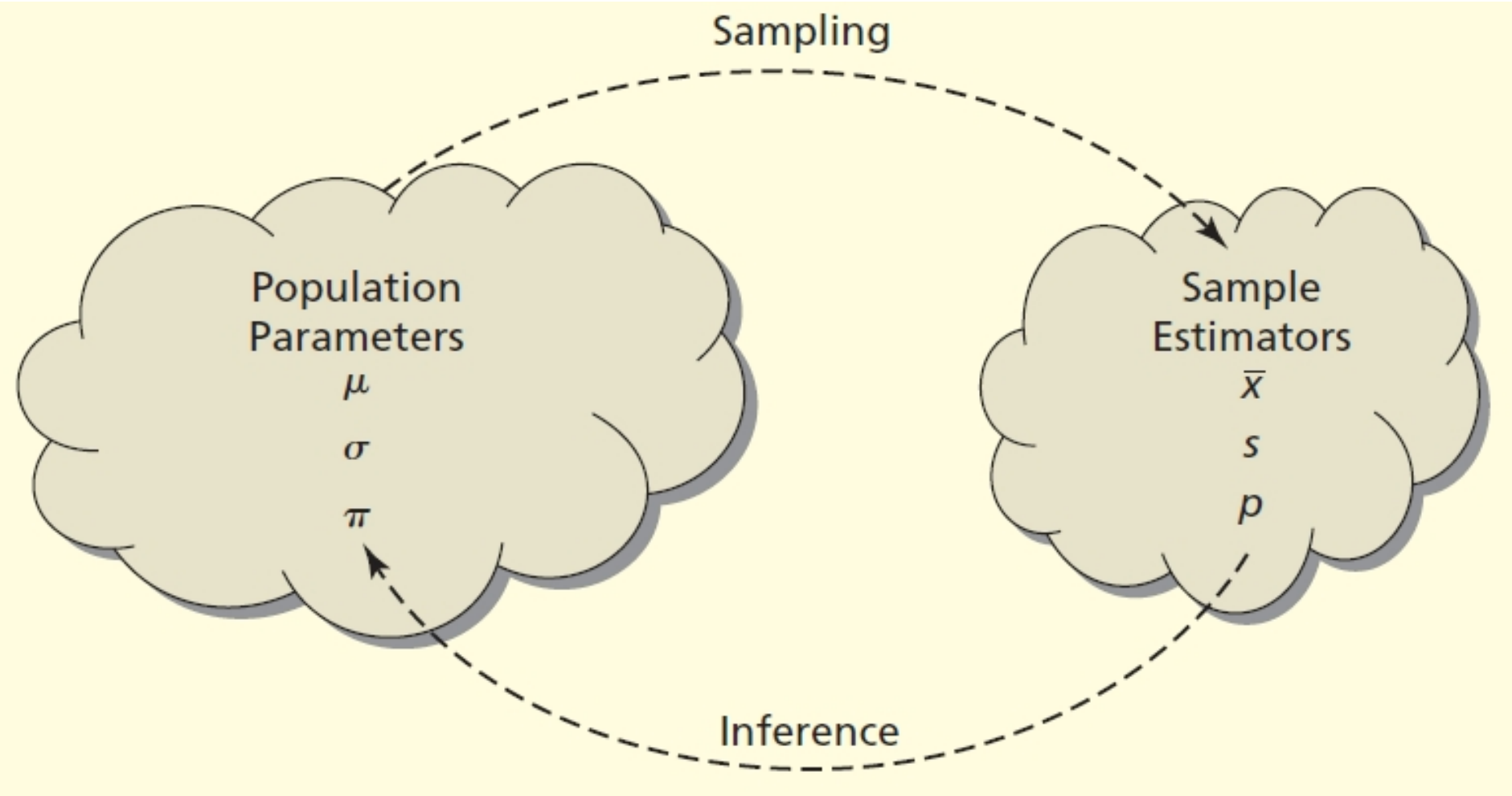
- Probability distributions
  - Normal distribution

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



- Functions for the normal distribution incluide `rnorm`, `dnorm`, `pnorm`, `qnorm`

Probability & Statistical Methods



Source: Doane and Seward (2011)

- Summary statistics
  - Sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Sample variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

    . . . and standard deviation $s = \sqrt{s^2}$

  - Median - middle value of a data set not sensitive to extreme values
    First sort the data

$$x_1 \leq x_2 \leq \cdots x_n$$

    If n is odd

$$m = x_{(n+1)/2}$$

    If n is even

$$m = \frac{x_{(n/2)} + x_{(n+2)+1}}{2}$$

- The sample mean we calculated above is a *point estimate*
- While this is informative it is important to understand the uncertainty in the estimate
- Confidence intervals do just that
- Suppose we wish to find a confidence interval for the mean of a distribution based on a sample

$$\bar{x} \pm Z_{\alpha/2} \cdot \sigma / \sqrt{n}$$

where α represents the *significance level*, Z is the quantile from the standard normal distribution, σ is the standard deviation, and n is the sample size

- The confidence level is $1 - \alpha$
- The confidence level is the probability that the interval includes the true mean

Example

A sample of size $n = 40$ is drawn from a population with a variance $\sigma^2 = 10$, and $\bar{x} = 7.164$. What is a 95% confidence interval for μ?

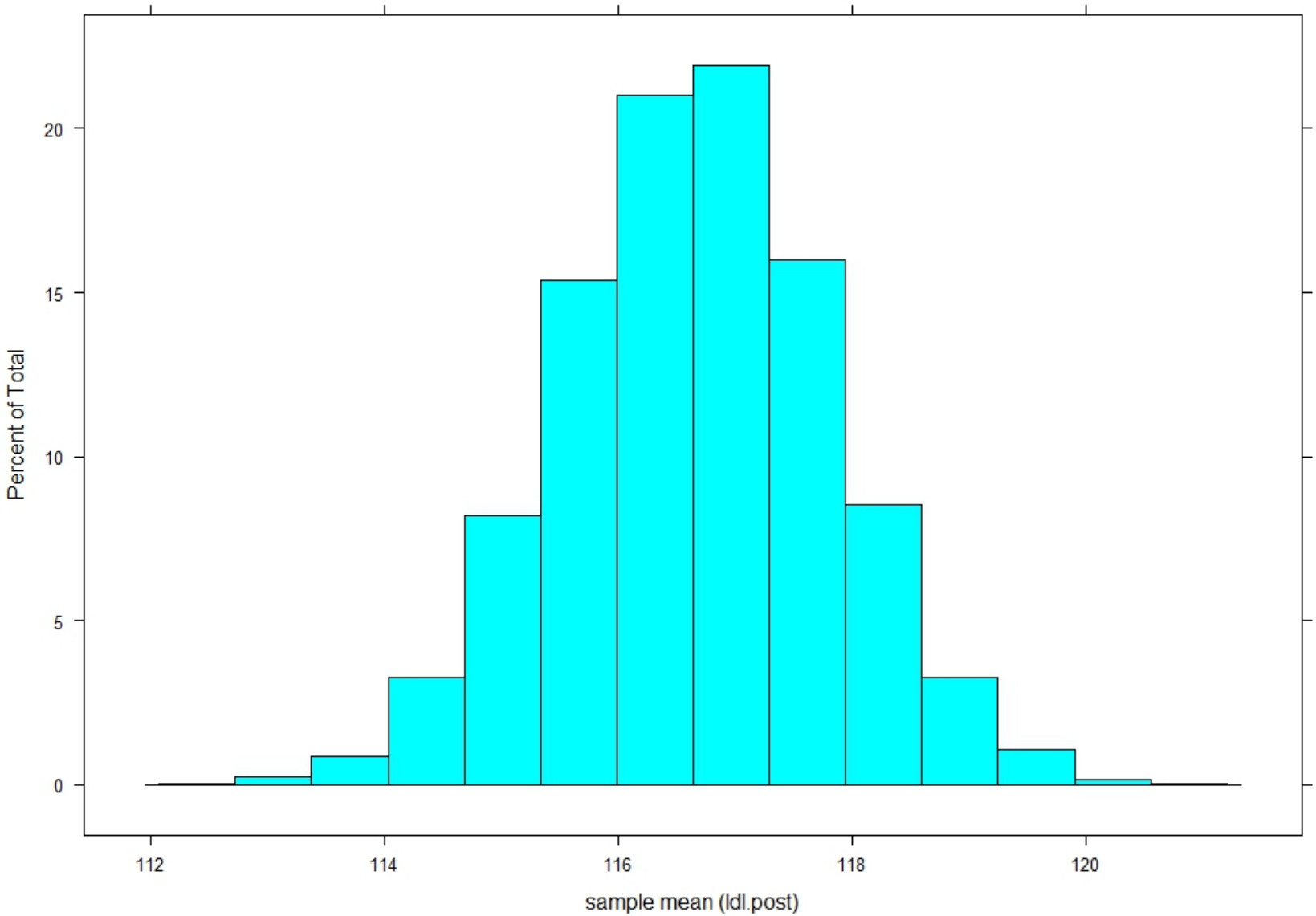
$$\left(7.164 - 1.96\sqrt{\frac{10}{40}}, \quad 7.164 + 1.96\sqrt{\frac{10}{40}}\right)$$

which is $(6.184, 8.144)$

- Sampling distributions
  The sampling distribution of an estimator is the probability distribution of all possible values the statistic may assume when a random sample of size n is taken
- The sampling distribution of the sample mean $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Try this . . .

```
library(lattice)
library(resample)
x <- ldl$ldl.post
samp.idx <- samp.bootstrap(n = 100, R = 10000, size =20)
mean(rowMeans(matrix(x[samp.idx], nrow = 10000, byrow = FALSE)))
mean(x)
histogram(rowMeans(matrix(x[samp.idx], nrow = 10000, byrow = FALSE)))
```


Percent of Total
0
5
10
15
20
112
114
116
118
120
sample mean (ldl.post)

- Covariance and correlation measure the association between two random variables given random variables X and Y

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] \\ &= \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]\end{aligned}$$

and

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}\left(X\right)\text{Var}\left(Y\right)}}$$

- While both measure association, correlation ρ is normalized and ranges from -1 to 1 making it easier to interpret

- We can also write correlation as

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y}$$

  when working with a sample, $\overline{X}$ is the sample mean for X, $\overline{Y}$ is the sample mean for Y, S_X and S_Y are the sample standard deviation for X and Y respectively, and n is the sample size
- In R we use the `cor` function for correlation and `cov` for covariance

```
library(ggplot2)
cor(economics$pce, economics$psavert)
[1] -0.837069
```

Probability & Statistical Methods

- Hypothesis testing is used to test assumptions and theories and guide our decisions
- State the level of significance α: the probability that you incorrectly reject H_0
- The levels of significance correspond to various quantiles of the standard normal distribution ...
  - $\alpha = 0.05, Z = 1.96,$
  - $\alpha = 0.10, Z = 1.645$
  - Refer to a Z table that provides the area under the curve for the cumulative normal distribution

- Steps in hypothesis testing:

Step 1: State the null and alternative hypotheses, H_0, and H_a (or H_1), determine if this is a one-sided or two-sided test

Step 2: State α, the level of significance

Step 3: Calculate the test statistic from the sample mean and variance

$$Z = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

Step 4: Compare the calculated test statistic to the critical value

Step 5: Accept H_0 if Z is in the acceptance region, or reject H_0 if Z is in the critical or rejection region

Example

Historically the test scores on a given math test are normally distributed with mean 75 and variance 36. A group of 16 students take the test this year, and score an average of 82. Do this group's math scores differ from the historical average? Test this at a 0.10 level of significance.

Solution

$$Z = \frac{82 - 75}{\sqrt{36/16}} = 4.67.$$

Since $\alpha = 0.1$ *we have a* $1 - \alpha$ *confidence level, and the critical value for* $Z = 1.645$.
Since $4.67 > 1.645$ *we reject* H_0 *and conclude that this year's math scores have a higher average than the historical average.*

Example

The specifications for a cable call for a mean breaking strength of 2000 pounds. For a sample of the cable the mean breaking strngth is 1955 pounds with a *standard error of the mean* of 25. Using $\alpha = 0.05$ determine if the difference in mean breaking strengths is statistically significant.

- In these examples the variance is known so we use the standard normal distribution
- If we don't know the variance and the sample size is small (*rule of thumb* $n < 30$) then use a t-distribution

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

- Compare this test statistic to the critical value from a t-table, for $n - 1$ degrees of freedom

Exercise

The copier at Seidenberg can make 45 copies per minute on average. The machine is altered to increase output, and three test runs produce 46, 47, and 48 copies per minute. Is this increase statistically significant? Use $\alpha = 0.05$. *Note* this is a one-tailed test.

Conduct a t-test in R using the function `t-test`

```
> t.test(ldl$ldl.post, mu = 117)

One Sample t-test

data:  ldl$ldl.post
t = -0.62329, df = 99, p-value = 0.5345
alternative hypothesis: true mean is not equal to 117
95 percent confidence interval:
115.6613 117.6987
sample estimates:
mean of x
116.68
```

We can do a two-sample t-test testing the equality of the sample means, and the null hypothesis is the difference is zero:
$H_0 : \mu_1 - \mu_2 = 0$

```
> t.test(ldl$ldl.post~ldl$gender)

Welch Two Sample t-test

data:  ldl$ldl.post by ldl$gender
t = 0.65255, df = 88.233, p-value = 0.5157
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.374449  2.718449
sample estimates:
mean in group f mean in group m
117.016         116.344
```

When testing for a difference in means, the test statistic

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\overline{X}_1 - \overline{X}_2}}$$

has a t-distribution when $n_1 + n_2 \leq 30$

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_i^2 = \frac{n_i \sum X_i^2 - (\sum X_i)^2}{n_i(n_i - 1)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n_i - 1}$$

For a given significance level, α, the critical value is $t_{\alpha/2,df}$ where $df = n_1 + n_2 - 2$

We accept H_0 if $-t_{\alpha/2,df} \leq t \leq t_{\alpha/2,df}$

Example

Two sets of IQ scores are sampled. For Group 1 $\overline{X} = 104, S = 10$ and $n = 16$; for group 2 $\overline{X} = 112, S = 8$ and $n = 14$. Is there a statistically significant difference in means between the two groups?

TABLE E.3

Critical Values of t

For a particular number of degrees of freedom, entry represents the critical value of t corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).

Degrees of Freedom	Cumulative Probabilities: 0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas: 0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564

- Arguably regression analysis is the most commonly used statistical model
- Linear regression explores relationships between variables that are linear
- Many problems can be analyzed using linear regression . . .
- And when the relationships are not quite linear, the transformation of the original variables results in linear relationships among the transformed variables
- The **functional** relation between two variables is

$$Y = f(X)$$

Example

Let's consider the relation between total dollar sales (Y) of a product and the number of units sold (X). If the unit price is \$2 then the functional relation is

$$Y = 2X.$$

Probability & Statistical Methods

- Unlike a functional relation a statistical relation is not perfect
- The observations do not fall on directly the curve
- Two key concepts of a statistical relationship captured by a regression model:
  1. The tendency of the dependent variable Y to vary with the independent variable X
  2. The points scatter around the curve of the statistical relationship
- The linear regression equation is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

for a simple linear regression and

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{1.1}$$

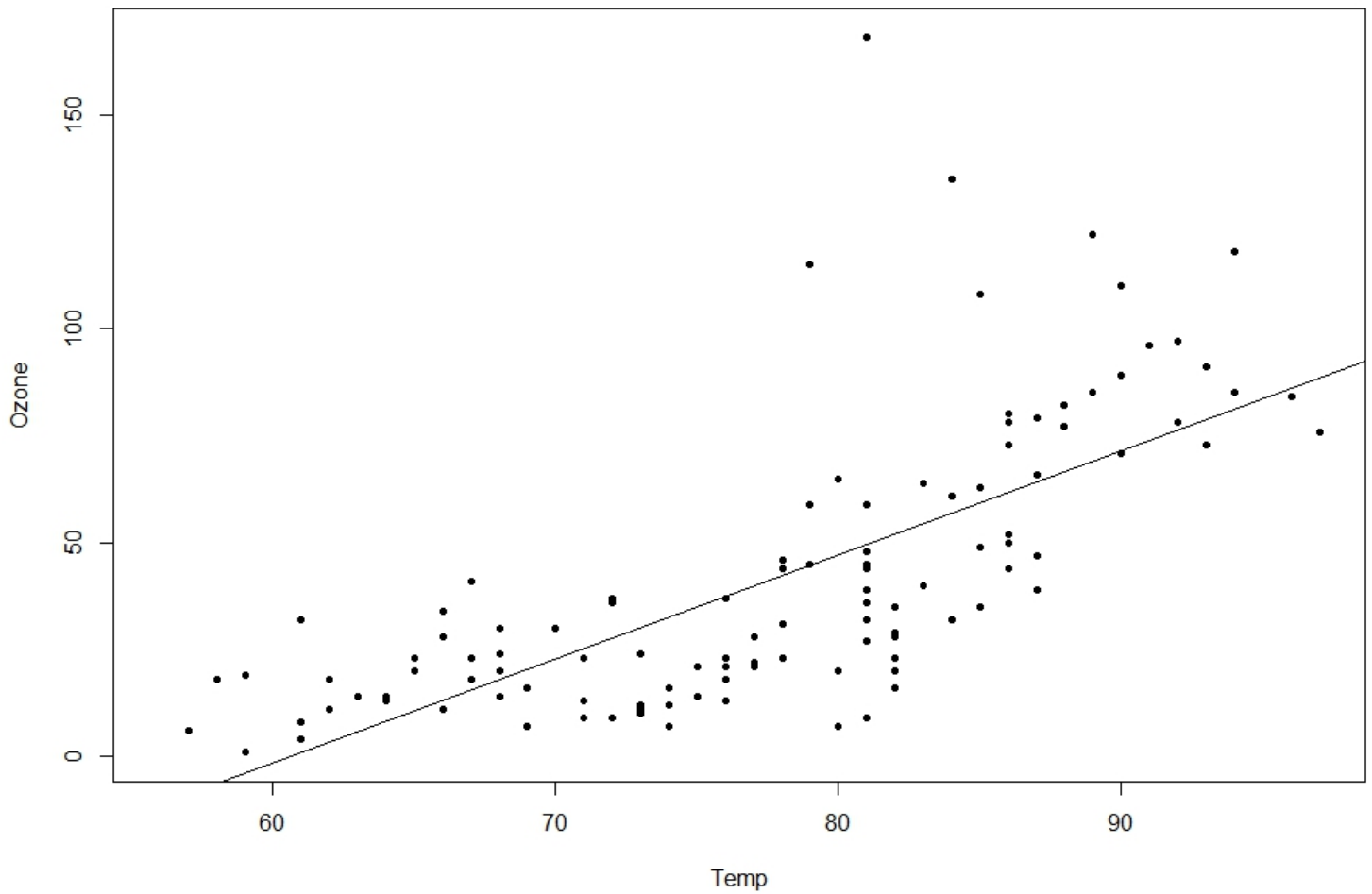
for a multiple regression model

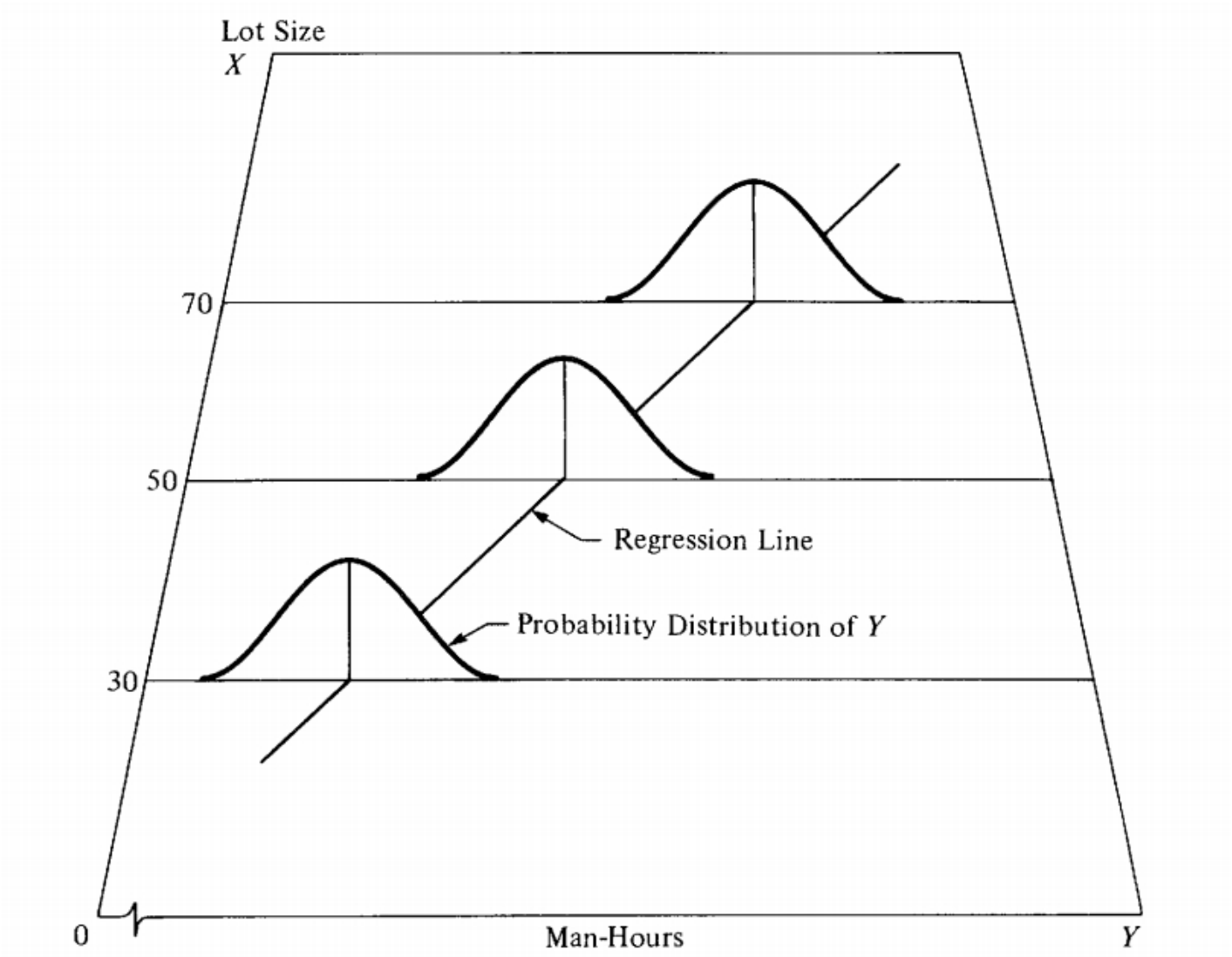
In (1.1) The β_i are the regression coefficients corresponding the the X_i independent variables, aka predictor variables, or explanatory variables, and ϵ is the random error, $\epsilon \sim N\left(0, \sigma^2\right)$

When we fit a regression model we estimate the coefficients and our model is denoted by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

The following graph shows the data points and the fitted regression line for `Ozone` regressed on `Temp` from the `airquality` data frame


Ozone
150
100
50
0
60
70
80
90
Temp



Neter et al. (1983)

Formulæfor $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Ordinary least squares: find the coefficients that minimize the sum of the squared error terms $\sum_{i=1}^{n}(Y_i - \hat{Y})^2$

NOTE: Let $\theta(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right]^2$. To minimize, take the partial derivatives, $\frac{\partial \theta(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0}$ and $\frac{\partial \theta(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1}$, set them to zero and solve for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Formal assumptions of regression analysis:

- The relation is, in fact, linear, so that the errors all have expected value $\mathbb{E}[\epsilon_i] = 0$ for all i
- The errors all have the same variance, $\text{Var}(\epsilon_i) = \sigma^2$ for all i
- The errors are independent of each other
- The errors are all normally distributed, $\epsilon_i \sim N(0, \sigma^2)$

The last two points are often stated as the errors are independent and identically distributed, or iid

Fitting a regression model in R

```
## Create a link to the website
AdLink <- "http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv"

## Use read.csv to get the data
Advertising <- read.csv(file = AdLink, sep = ",", header = TRUE)

## Fit the regression
ad.lm <- lm(Sales~TV, data = Advertising)

## View the results
summary(ad.lm)
```

SOURCE: The data are taken from ”An Introduction to Statistical Learning, with applications in R” (Springer, 2013)

- Analysis of Variance (ANOVA)
  - Used in experimental situations
  - Apply several treatments or treatment combinations to randomly selected experimental units
  - Compare the treatment means for some response y

Example

Suppose that a researcher has developed two chemical additives for increasing miles per gallon in cars. Without additives a gallon yields an average of μ miles. If chemical 1 is added, the mileage is expected to increase by τ_1 miles per gallon, and if chemical 2 is added, the mileage would increase by τ_2 miles per gallon.

- To test the effects of these additives we need more than one car for each type of additive
- Suppose we have three cars for each type our ANOVA model is

$$y_{11} = \mu + \tau_1 + \epsilon_{11}, \quad y_{12} = \mu + \tau_1 + \epsilon_{12}, \quad y_{13} = \mu + \tau_1 + \epsilon_{13}$$
$$y_{21} = \mu + \tau_2 + \epsilon_{21}, \quad y_{22} = \mu + \tau_2 + \epsilon_{22}, \quad y_{23} = \mu + \tau_2 + \epsilon_{23}$$

or

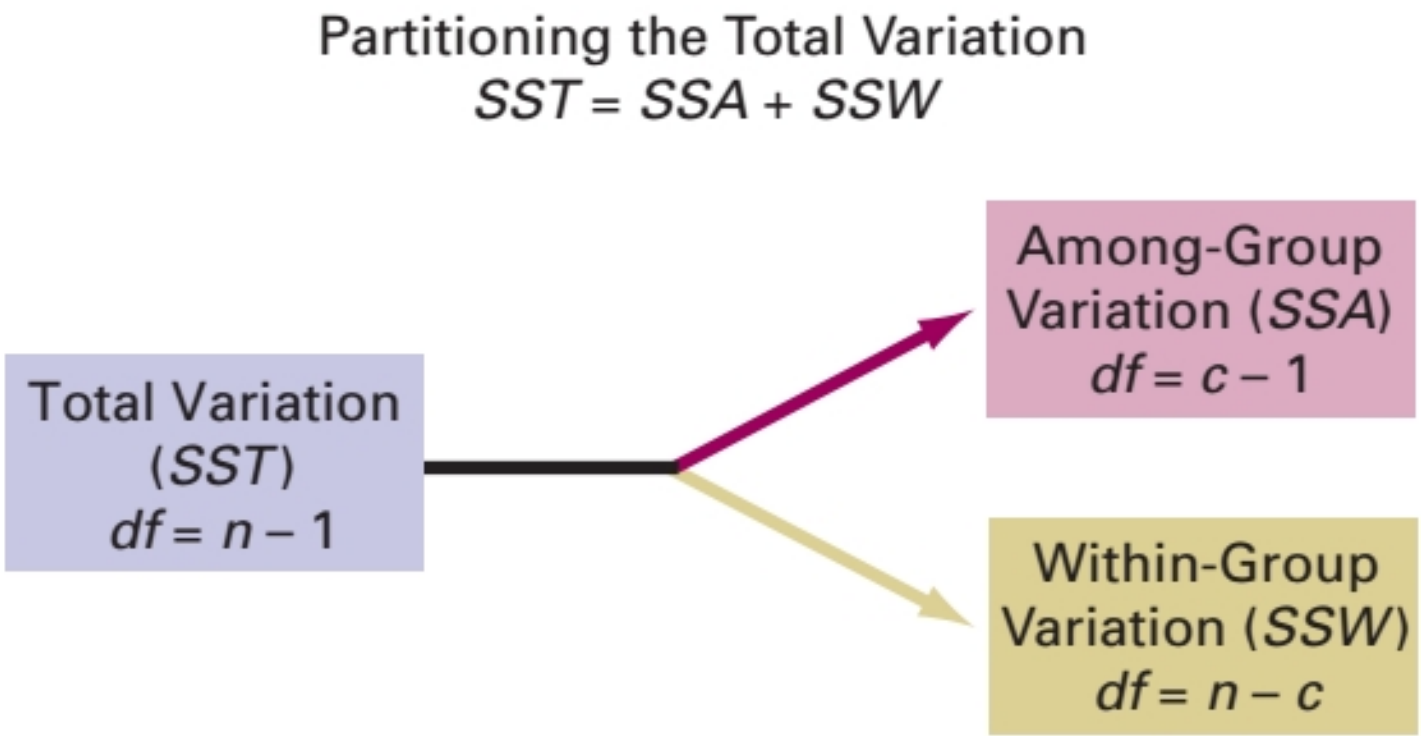
$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

and we test if the average y for treatment 1 is the same as the average y for treatment 2

Assumptions for Analysis of Variance

- The response variable is normally distributed for all the groups
- The variance of the response variable, denoted σ^2, is the same for all the groups
- The observations are independent

- *Analysis of Variance* is misleading because the objective in ANOVA is to analyze differences among the group means, not the variances
- By analyzing the variation among and within the groups, you can reach conclusions about possible differences in group means
- Total variation is subdivided into variation that is due to differences among the groups and variation that is due to differences within the groups
- Within-group variation measures random variation
- Among-group variation is due to differences from group to group
- The symbol c (in this text) is used to indicate the number of groups


Partitioning the Total Variation
SST = SSA + SSW
Total Variation (SST) df = n – 1
Among-Group Variation (SSA) df = c – 1
Within-Group Variation (SSW) df = n – c

Total Variation

$$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \overline{\overline{X}})^2 \tag{1.2}$$

where

$$\overline{\overline{X}} = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}}{n} = \text{grand mean}$$

$$X_{ij} = i\text{th value in group } j$$

$$n_j = \text{number of values in group } j$$

$$n = \text{total number of observations (all groups)}$$

$$c = \text{number of groups}$$

Among-Group Variation

$$SSA = \sum_{j=1}^{c} n_j(\overline{X}_j - \overline{\overline{X}})^2 \tag{1.3}$$

where

$$\begin{aligned} c &= \text{number of groups} \\ n_j &= \text{number of values in group } j \\ \overline{X}_j &= \text{sample mean of group } j \\ \overline{\overline{X}} &= \text{grand mean} \end{aligned}$$

Within-Group Variation

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2 \tag{1.4}$$

where

$$c = \text{number of groups}$$
$$X_{ij} = i\text{th value in group } j$$
$$\overline{X}_j = \text{sample mean of group } j$$

To test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$$

against the alternative

$$H_1 : \text{Not all } \mu_j \text{ areequal}$$

we create an ANOVA summary table

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F	p-value
Among-groups	$c-1$	SSA	$MSA = \frac{SSA}{c-1}$	$F = \frac{MSA}{MSW}$	$\Pr(F > F^*)$
Within-groups	$n-c$	SSW	$MSW = \frac{SSW}{n-c}$		
Total	$n-1$	SST			

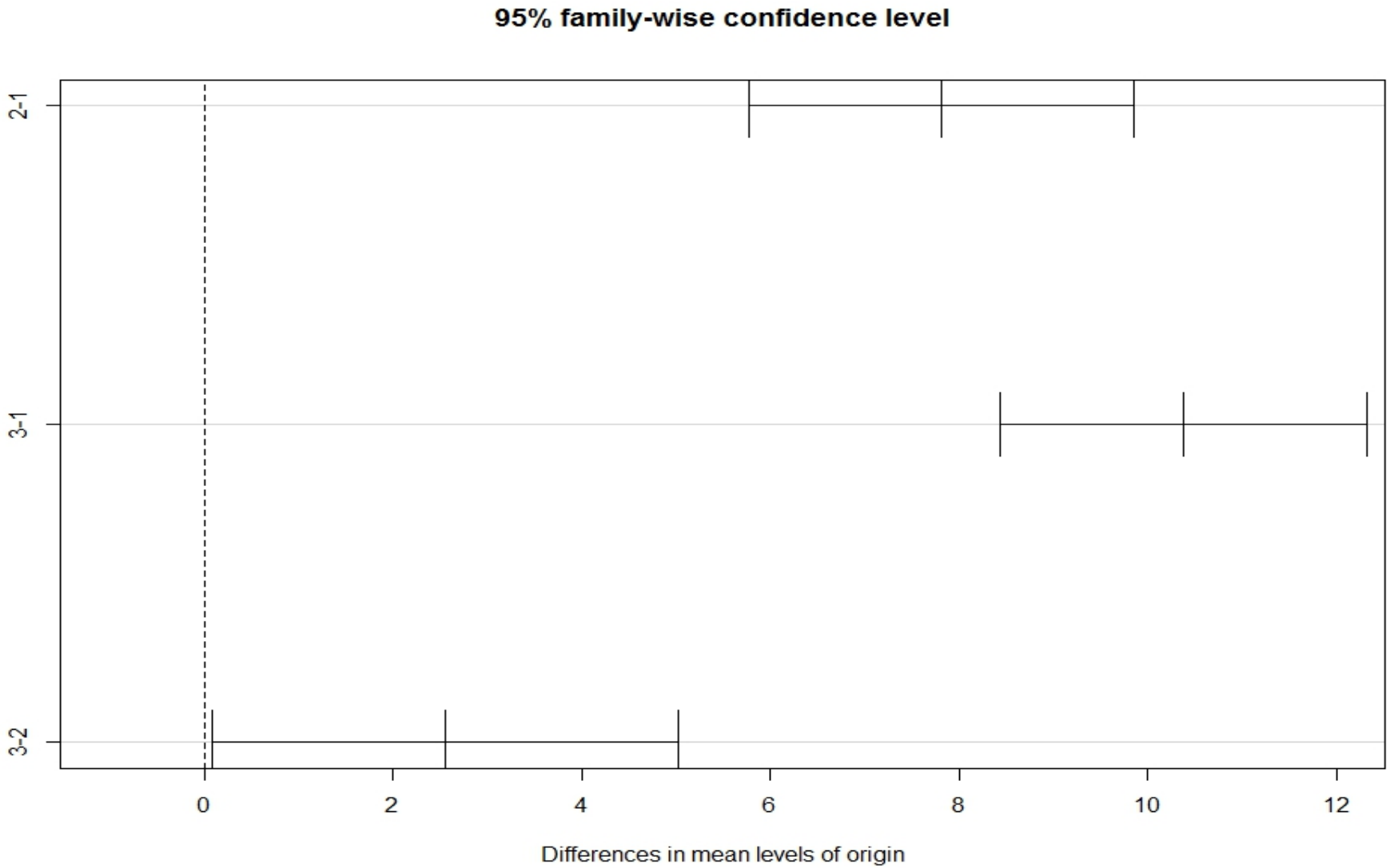
Fitting an ANOVA model in R

```
## Create a link to the website
AutoLink <- "http://www-bcf.usc.edu/~gareth/ISL/Auto.csv"

## Use read.csv to get the data
Auto <- read.csv(AutoLink, sep = ",", header = TRUE)

## Make origin a factor
Auto["origin"] <- lapply(Auto["origin"], factor)
# OR
Auto$origin <- factor(Auto$origin)

## Fit the ANOVA model
auto.aov <- aov(mpg~origin, data = Auto)

## View the results
summary(auto.aov)
```

SOURCE: The data are taken from ”An Introduction to Statistical Learning, with applications in R” (Springer, 2013)

- `summary(auto.aov)` shows us there is a significant difference in the means across origins but doesn't tell us what the differences are
- Tukey Honest Significant Differences provides us with that information

```
TukeyHSD(auto.aov, ordered = TRUE)
plot(TukeyHSD(auto.aov, "origin"), xlim = c(-1,12))
```


95% family-wise confidence level
2-1
3-1
3-2
0
2
4
6
8
10
12
Differences in mean levels of origin

Probability & Statistical Methods

- χ^2 test of independence
  - A test of independence for two categorical variables
  - Create a contingency table that has r rows and c columns
  - H_0 : The two categorical variables are independent
  - H_1 : The two categorical variables are dependent
  - Compute the test statistic

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

  - Reject H_0 if $\chi^2 > \chi^2_\alpha$, the critical value with $(r-1)(c-1)$ df

- Non-parametric statistics: Wilcoxon Rank Sum Test
  - A nonparametric procedure does not depend on the assumption of normality for the two populations
  - Use when the sample sizes are small and you cannot assume that the data in each sample are from normally distributed populations

Calculating the χ^2 statistic

f_o = observed values in the cell
f_e = expected values in a cell, calculated as follows:

$$f_e = \frac{\text{row total} \times \text{column total}}{n}$$

	Golden Palm	Palm Royale	Palm Princess
Price	23	7	37
Location	39	13	8
Accommodation	13	5	13
Other	13	8	8

Enter the numerical data in a text file separated by commas, and execute the following code

```
chiexam <- read.csv("chidata.txt", header = FALSE)
names(chiexam) <- c("Golden Palm","Palm Royale","Palm Princess")
rownames(chiexam) <- c("Price", "Location", "Accommodation", "Other")
chisq.test(chiexam)
```

```
> chisq.test(chiexam)

Pearson's Chi-squared test

data:  chiexam
X-squared = 27.41, df = 6, p-value = 0.000121
```

	Annual Storm Count		
Wind Speed (km/h)	1	2	3 or more
$119 \leq W_i < 154$	13	12	28
$154 \leq W_i < 179$	2	6	12
$179 \leq W_i$	12	6	17

Conduct a χ^2-test to determine if wind speed and annual storm count are independent

Source: Parisi and Lund (2000)

- The Wilcoxon rank sum test tests whether there is a difference in the medians from two samples
- If the two sample sizes are unequal, n_1 represents the smaller sample and n_2 the larger sample
- The Wilcoxon rank sum test statistic, T_1, is defined as the sum of the ranks assigned to the n_1 values in the smaller sample
- If T_2 is the sum of the ranks assigned to the n_2 items in the second sample, then

$$T_1 + T_2 = \frac{n(n+1)}{2}$$

  which follows from the well known result from mathematical induction
- The prior point checks for the accuracy in assigning the ranks

When the sample sizes are large the test statistic T_1 is approximately normally distributed with

$$\mu_{T_1} = \frac{n_1(n+1)}{2}$$

and

$$\sigma_{T_1} = \sqrt{\frac{n_1 n_2 (n+1)}{12}}$$

The test statistics is

$$Z = \frac{T_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}$$

which approximately follows a standard normal distribution

- Steps for performing the Wilcoxon rank sum test:
  - Replace the values in the two samples of size n_1 and n_2 with their combined ranks (unless the data are the ranks)
  - Define $n = n_1 + n_2$ as the total sample size
  - Assign the ranks so that rank 1 is given to the smallest of the n combined values, rank 2 is given to the second smallest, and so on
  - For ties, assign each value the average of the ranks that would have been assigned had there been no ties
  - Compute T_1 and Z
  - Compare Z to the critical value and accept or reject H_0 as appropriate

Table: Time to second heart attack for smokers and nonsmokers

Smoker	NonSmoker
1.09	1.96
3.50	0.07
3.69	0.01
0.63	10.21
0.35	0.10
3.34	0.25
1.93	0.41
1.57	3.26
1.79	3.57
0.86	0.24
0.03	0.51
0.19	0.33
0.98	7.02
0.26	5.85
1.01	1.83

When n_1 and n_2 are both ≤ 10 you use the table of lower and upper crittical values

	α		n_1						
n_2	One-tail	Two-tail	4	5	6	7	8	9	10
4	0.05	0.10	11,25						
	0.025	0.05	10,26						
	0.01	0.02	—,—						
	0.005	0.01	—,—						
5	0.05	0.10	12,28	19,36					
	0.025	0.05	11,29	17,38					
	0.01	0.02	10,30	16,39					
	0.005	0.01	—,—	15,40					
6	0.05	0.10	13,31	20,40	28,50				
	0.025	0.05	12,32	18,42	26,52				
	0.01	0.02	11,33	17,43	24,54				
	0.005	0.01	10,34	16,44	23,55				
7	0.05	0.10	14,34	21,44	29,55	39,66			
	0.025	0.05	13,35	20,45	27,57	36,69			
	0.01	0.02	11,37	18,47	25,59	34,71			
	0.005	0.01	10,38	16,49	24,60	32,73			
8	0.05	0.10	15,37	23,47	31,59	41,71	51,85		
	0.025	0.05	14,38	21,49	29,61	38,74	49,87		
	0.01	0.02	12,40	19,51	27,63	35,77	45,91		
	0.005	0.01	11,41	17,53	25,65	34,78	43,93		
9	0.05	0.10	16,40	24,51	33,63	43,76	54,90	66,105	
	0.025	0.05	14,42	22,53	31,65	40,79	51,93	62,109	
	0.01	0.02	13,43	20,55	28,68	37,82	47,97	59,112	
	0.005	0.01	11,45	18,57	26,70	35,84	45,99	56,115	
10	0.05	0.10	17,43	26,54	35,67	45,81	56,96	69,111	82,128
	0.025	0.05	15,45	23,57	32,70	42,84	53,99	65,115	78,132
	0.01	0.02	13,47	21,59	29,73	39,87	49,103	61,119	74,136
	0.005	0.01	12,48	19,61	27,75	37,89	47,105	58,122	71,139

References

Doane, D. P. and Seward, L. E. (2011).
Applied Statistics in Business and Economics.
McGraw-Hill/Irwin, third edition.

Kabacoff, R. I. (2015).
R in Action.
Manning, Shelter Island, NY, second edition.

Lander, J. P. (2014).
R for Everyone.
Addison-Wesley, Upper Saddle River.

Neter, J., Wasserman, W. and Kutner, M. H. (1983).
Applied LInear Regression Models.
Richard D. Irwin, Inc, Homewood, IL.

Parisi, F. and Lund, R. (2000).
Seasonality and return periods of landfalling Atlantic basin hurricanes.
Australian & New Zealand Journal of Statistics **42**, 271–282.

References (cont.)

Zumel, N. and Mount, J. (2014).
Practical Data Science with R.
Manning, Shelter Island, NY, second edition.