David S. Moore • George P. McCabe

Introduction to the Practice of Statistics

Fifth Edition

Chapter 10:

Inference for Regression

Copyright © 2005 by W. H. Freeman and Company

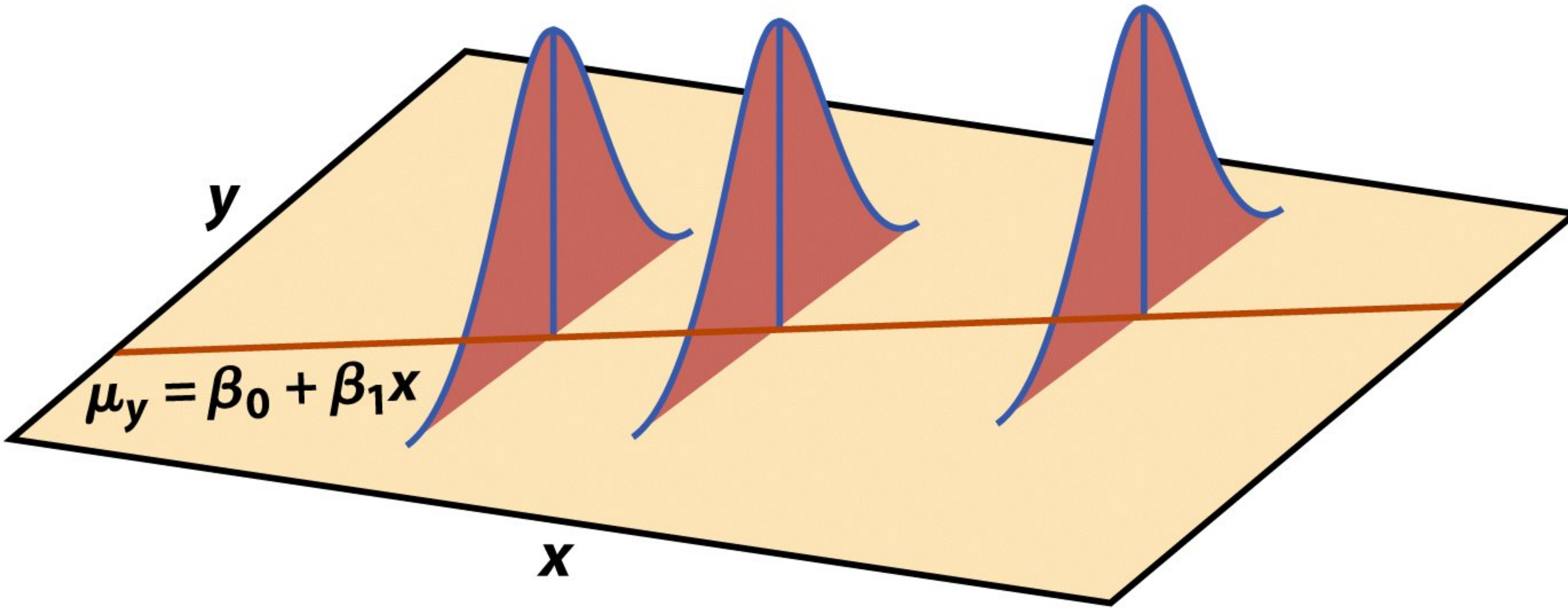


Figure 10-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

SIMPLE LINEAR REGRESSION MODEL

Given n observations on the explanatory variable x and the response variable y,

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Here $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviations ϵ_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ.

The parameters of the model are β_0, β_1, and σ.

Definition, pg 638
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

```
The regression equation is
MPG = - 7.80 + 7.87 logmph

Predictor        Coef       StDev        T        P
Constant       -7.796       1.155    -6.75    0.000
logmph         7.8742      0.3541    22.24    0.000

S = 0.9995      R-Sq = 89.5%     R-Sq(adj) = 89.3%
```

Figure 10-5b

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

Root MSE	0.99952	R-Square	0.8950
Dependent Mean	17.72500	Adj R-Sq	0.8932
Coeff Var	5.63902		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	–7.79625	1.15494	–6.75	<.0001	–10.10812	–5.48438
logmph	1	7.87422	0.35411	22.24	<.0001	7.16539	8.58305

Figure 10-5e
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR REGRESSION SLOPE AND INTERCEPT

A **level C confidence interval for the intercept** β_0 is

$$b_0 \pm t^* \mathrm{SE}_{b_0}$$

A **level C confidence interval for the slope** β_1 is

$$b_1 \pm t^* \mathrm{SE}_{b_1}$$

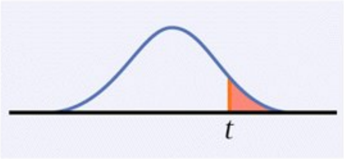
In these expressions t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

To test the hypothesis $H_0: \beta_1 = 0$, compute the **test statistic**

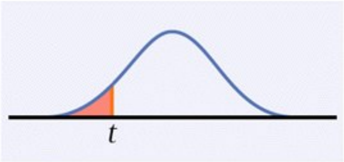
$$t = \frac{b_1}{\mathrm{SE}_{b_1}}$$

The **degrees of freedom** are $n - 2$. In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

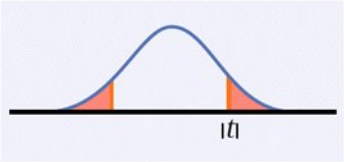
$H_a: \beta_1 > 0$ is $P(T \geq t)$



$H_a: \beta_1 < 0$ is $P(T \leq t)$



$H_a: \beta_1 \neq 0$ is $2P(T \geq |t|)$



Definition, pg 644
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

CONFIDENCE INTERVAL FOR A MEAN RESPONSE

A **level C confidence interval for the mean response** μ_y when x takes the value x^* is

$$\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Definition, pg 647
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

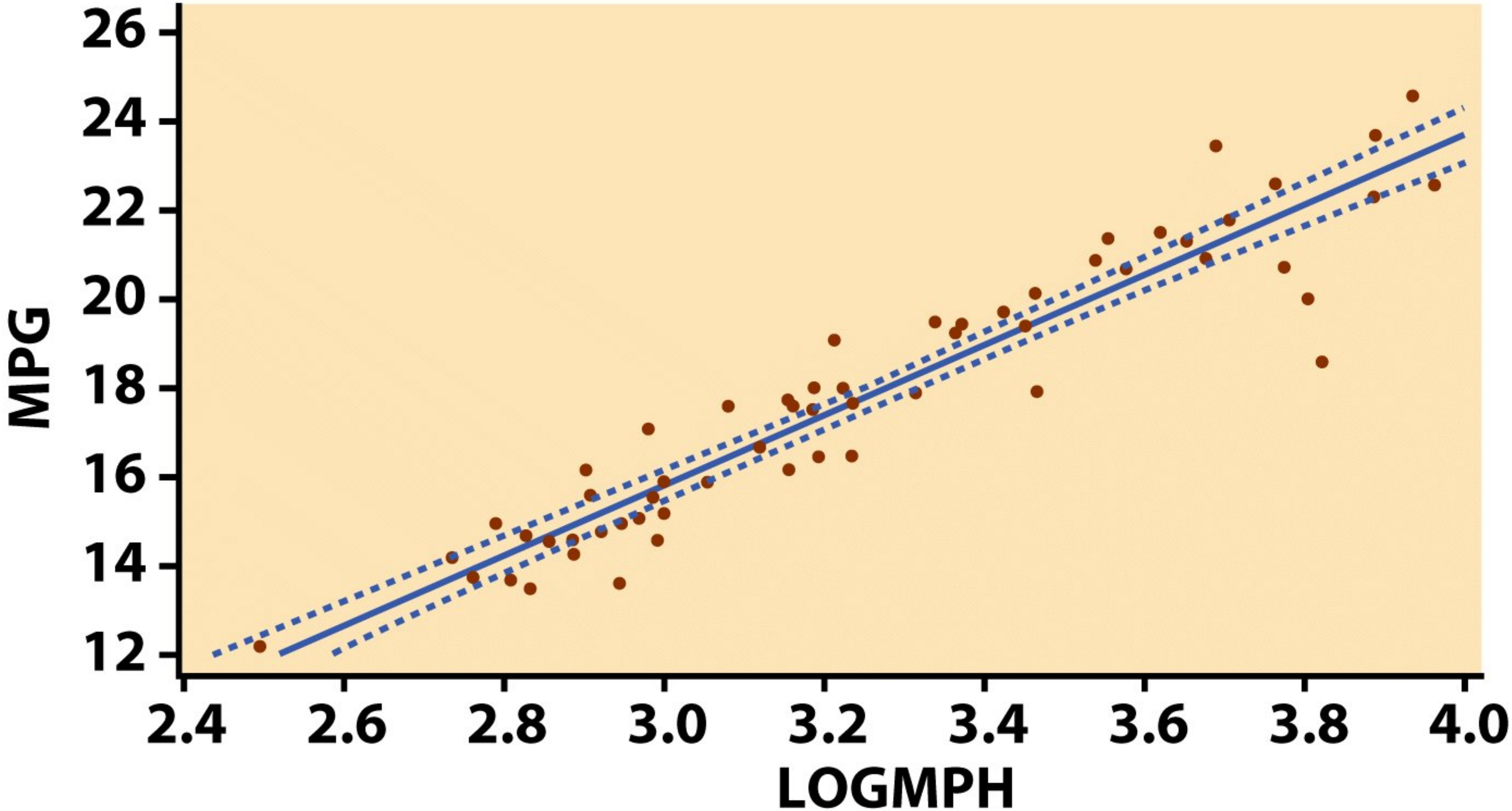


Figure 10-9
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

PREDICTION INTERVAL FOR A FUTURE OBSERVATION

A **level C prediction interval for a future observation** on the response variable y from the subpopulation corresponding to x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Definition, pg 649
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

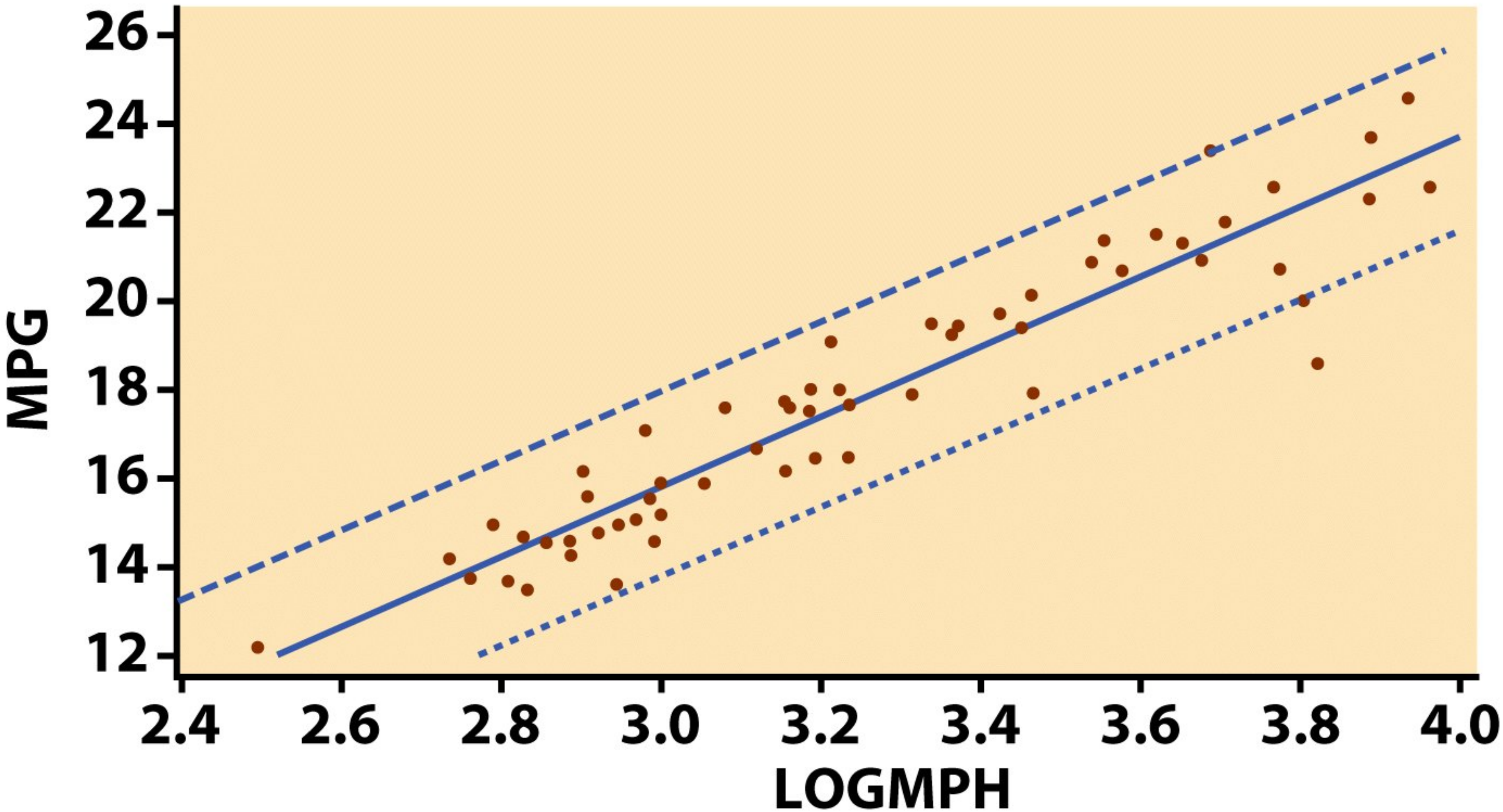


Figure 10-10
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

SUMS OF SQUARES, DEGREES OF FREEDOM, AND MEAN SQUARES

Sums of squares represent variation present in the responses. They are calculated by summing squared deviations. **Analysis of variance** partitions the total variation between two sources.

The sums of squares are related by the formula

$$\text{SST} = \text{SSM} + \text{SSE}$$

That is, the total variation is partitioned into two parts, one due to the model and one due to deviations from the model.

Degrees of freedom are associated with each sum of squares. They are related in the same way:

$$\text{DFT} = \text{DFM} + \text{DFE}$$

To calculate **mean squares,** use the formula

$$\text{MS} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

Definition, pg 654
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

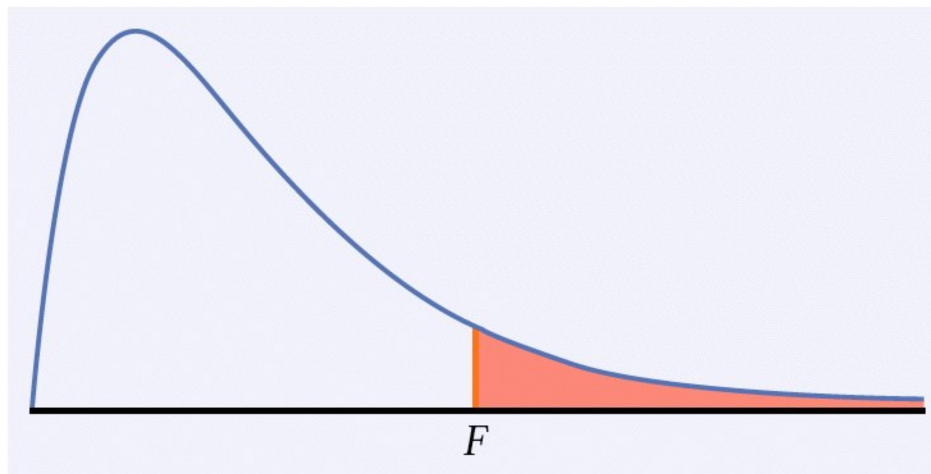
ANALYSIS OF VARIANCE *F* TEST

In the simple linear regression model, the hypotheses

$$H_0\colon \beta_1 = 0$$
$$H_a\colon \beta_1 \neq 0$$

are tested by the ***F* statistic**

$$F = \frac{\text{MSM}}{\text{MSE}}$$



The P-value is the probability that a random variable having the $F(1, n-2)$ distribution is greater than or equal to the calculated value of the F statistic.

Definition, pg 655
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	493.989	1	493.989	494.467	.000[a]
	Residual	57.944	58	.999		
	Total	551.932	59			

a. Predictors: (Constant), LOGMPH

b. Dependent Variable: MPG

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.946[a]	.895	.893	.9995

a. Predictors: (Constant), LOGMPH

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-7.796	1.155		-6.750	.000
	LOGMPH	7.874	.354	.946	22.237	.000

a. Dependent Variable: MPG

Figure 10-12

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

STANDARD ERRORS FOR ESTIMATED REGRESSION COEFFICIENTS

The standard error of the slope b_1 of the least-squares regression line is

$$\mathrm{SE}_{b_1} = \frac{s}{\sqrt{\sum(x_i - \overline{x})^2}}$$

The standard error of the intercept b_0 is

$$\mathrm{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}}$$

Definition, pg 660

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

STANDARD ERRORS FOR $\hat{\mu}$ AND $\hat{y}$

The standard error of $\hat{\mu}$ is

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2}}$$

The standard error for predicting an individual response $\hat{y}$ is[4]

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2}}$$

Definition, pg 662

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

TEST FOR A ZERO POPULATION CORRELATION

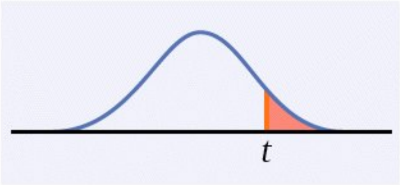
To test the hypothesis H_0: $\rho = 0$, compute the t statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

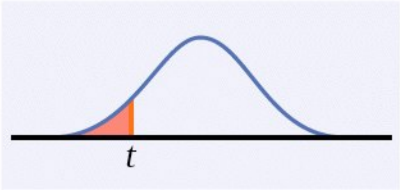
where n is the sample size and r is the sample correlation.

In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

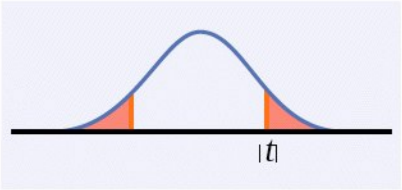
H_a: $\rho > 0$ is $P(T \geq t)$



H_a: $\rho < 0$ is $P(T \leq t)$



H_a: $\rho \neq 0$ is $2P(T \geq |t|)$



Definition, pg 664

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

Correlations

		LOGMPH	MPG
LOGMPH	Pearson Correlation	1	.946**
	Sig. (2-tailed)	.	.000
	N	60	60
MPG	Pearson Correlation	.946**	1
	Sig. (2-tailed)	.000	.
	N	60	60

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 10-14
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company