STA 4210 – Supplementary Notes and R Programs

Larry Winner
Department of Statistics
University Of Florida

Introduction/Review

Mathematical Operations – Summation Operators

Consider sequences of numbers and numeric constants.

Sum of a sequence of Variables: $\sum_{i=1}^{n} Y_i = Y_1 + ... + Y_n$

Sum of a sequence of Constants: $\sum_{i=1}^{n} k = k + ... + k = nk$

Sum of a sequence of Sums of Variables: $\sum_{i=1}^{n} \left(X_i + Z_i \right) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Z_i$

Sum of a sequence of (Commonly) Linearly Transformed Variables: $\sum_{i=1}^{n} \left(a + bX_i \right) = na + b\sum_{i=1}^{n} X_i$

Sum of a sequence of (Individually) Linearly Transformed Variables: $\sum_{i=1}^{n} \left(a_i + b_i X_i \right) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i X_i$

Sum of a sequence of Sums of Multiples of Variables: $\sum_{i=1}^{n} \left(a_i X_i + b_i Z_i \right) = \sum_{i=1}^{n} a_i X_i + \sum_{i=1}^{n} b_i Z_i$

Example – Opening Weekend Box-Office Gross for Harry Potter Films

Date	Movie	Gross($M)	Theaters	PerTheater($K)	Euros/Dollar	Gross (€M)
11/16/2001	Sorcerer's Stone	90.29	3672	24.59	1.1336	102.36
11/15/2002	Chamber of Secrets	88.36	3682	24.00	0.9956	87.97
6/4/2004	Prisoner of Azkaban	93.69	3855	24.30	0.8135	76.21
11/18/2005	Goblet of Fire	102.69	3858	26.62	0.8496	87.24
7/13/2007	Order of the Phoenix	77.11	4285	18.00	0.7263	56.00
7/17/2009	Half-Blood Prince	77.84	4325	18.00	0.7085	55.15
11/19/2010	Deathly Hallows: Part I	125.02	4125	30.31	0.7353	91.93
7/15/2011	Deathly Hallows: Part II	169.19	4375	38.67	0.7042	119.14
Total		824.18	32,177.00			676.00

Total Gross ($Millions): $\sum_{i=1}^{n} Y_i = 90.29 + 88.36 + ... + 169.19 = 824.18$

Total Gross (Millions of Euros): $\sum_{i=1}^{n} a_i Y_i = 1.1336(90.29) + 0.9956(88.36) + ... + 0.7042(169.19) = 676.00$

Question: What is the average gross per theater for all movies? Is it the same as the average of individual movies per theater?

Basic Probability

Addition Theorem

A_i, A_j are 2 events defined on a sample space.

$P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j)$ where:

$P(A_i \cup A_j) \equiv$ Probability at least one occurs $P(A_i \cap A_j) \equiv$ Probability both occur

Multiplication Theorem (Can be obtained from counts when data are in contingency table)

$$P(A_i \mid A_j) = \frac{P(A_i \cap A_j)}{P(A_j)}$$ where $P(A_i \mid A_j) \equiv$ Probabilty A_i occurs given A_j has occured

$$P(A_j \mid A_i) = \frac{P(A_i \cap A_j)}{P(A_i)}$$

$$\Rightarrow P(A_i \cap A_j) = P(A_i) P(A_j \mid A_i) = P(A_j) P(A_i \mid A_j)$$

Complementary Events

$P(\overline{A}_i) = 1 - P(A_i)$ where $\overline{A}_i \equiv$ event A_i does not occur

$P(\overline{A_i \cup A_j}) = P(\overline{A}_i \cap \overline{A}_j)$

Example – New York City Sidewalk Cafes

Cafes classified by size (<100 ft^2, 100-199, 200-299, 300-399, 400-499, 500-599, ≥600) and type (enclosed, unenclosed).

Type\Size	<100	100-199	200-299	300-399	400-499	500-599	≥600	Total
Enclosed	2	18	31	30	23	7	9	120
Unenclosed	98	318	200	118	63	26	40	863
Total	100	336	231	148	86	33	49	983

Let $A_1 \equiv$ Size < 300ft^2 and $A_2 \equiv$ Type = Unenclosed.

$$P(A_1) = \frac{100+336+231}{983} = \frac{667}{983} = 0.6785$$

$$P(A_2) = \frac{863}{983} = 0.8779$$

$$P(A_1 \cap A_2) = P(A_1 A_2) = \frac{98+318+200}{983} = \frac{616}{983} = 0.6267$$

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 A_2) = \frac{667+863-616}{983} = \frac{914}{983} = 0.9298 = 0.6785 + 0.8779 - 0.6267$$

$$P(A_1 \mid A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} = \frac{616}{863} = 0.7138 = \frac{0.6267}{0.8779}$$

$$P(A_2 \mid A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{616}{667} = 0.9235 = \frac{0.6267}{0.6785}$$

$$P(\overline{A_1}) = \frac{148+86+33+49}{983} = \frac{316}{983} = 0.3215 = 1 - 0.6785$$

$$P(\overline{A_2}) = \frac{120}{983} = 0.1221 = 1 - 0.8779$$

$$P(\overline{A_1 \cup A_2}) = \frac{30+23+7+9}{983} = \frac{69}{983} = 0.0702 = P(\overline{A_1} \cap \overline{A_2})$$

Univariate Random Variables

Probability (Density) Functions

Discrete (RV $\equiv Y$ takes on masses of probability at specific points $Y_1,...,Y_k$):

$f(Y_s) = P(Y = Y_s) \quad s = 1,...,k$ often written $f(y)$ where y is specific point Y_s

Continuous (RV $\equiv Y$ takes on density of probability over ranges of points on continuum)

$f(Y) \equiv$ density at Y (confusing notation, often written $f(y)$ where y is specific point and Y is RV)

Expected Value (Long Run Average Outcome, aka Mean)

Discrete: $\mu_Y = E\{Y\} = \sum_{s=1}^{k} Y_s f(Y_s)$ Continuous: $\mu_Y = E\{Y\} = \int_{-\infty}^{\infty} Yf(Y)\,dY = \int_{-\infty}^{\infty} yf(y)\,dy$

a, c constants $\Rightarrow E\{a + cY\} = a + cE\{Y\} = a + c\mu_Y \Rightarrow E\{a\} = a \Rightarrow E\{cY\} = cE\{Y\} = c\mu_Y$

Variance (Average Squared Distance from Expected Value)

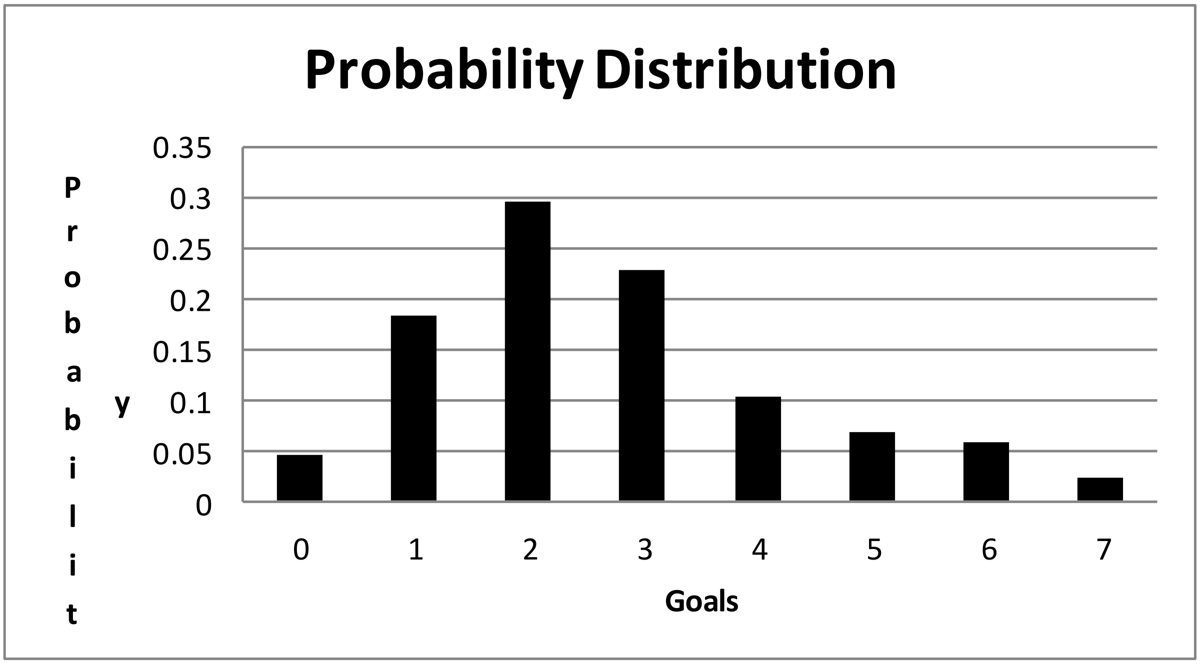
$$\sigma_Y^2 = \sigma^2\{Y\} = E\left\{(Y - E\{Y\})^2\right\} = E\left\{(Y - \mu_Y)^2\right\}$$

Equivalently (Computationally easier): $\sigma_Y^2 = \sigma^2\{Y\} = E\{Y^2\} - (E\{Y\})^2 = E\{Y^2\} - \mu_Y^2$

a, c constants $\Rightarrow \sigma^2\{a + cY\} = c^2\sigma^2\{Y\} = c^2\sigma_Y^2 \Rightarrow \sigma^2\{a\} = 0 \Rightarrow \sigma^2\{cY\} = c^2\sigma^2\{Y\} = c^2\sigma_Y^2$

<u>Example – Total Goals per Game in National Women's Soccer League Games (2013)</u>

Goals (y)	Frequency	Probability=p(y)	y*p(y)	(y^2)*p(y)
0	4	0.0455	0.0000	0.0000
1	16	0.1818	0.1818	0.1818
2	26	0.2955	0.5909	1.1818
3	20	0.2273	0.6818	2.0455
4	9	0.1023	0.4091	1.6364
5	6	0.0682	0.3409	1.7045
6	5	0.0568	0.3409	2.0455
7	2	0.0227	0.1591	1.1136
Total	88	1	2.7045	9.9091



Note: Using more common notation, where *y* represents a specific outcome (number of goals) and *p*(*y*) represents the probability of a game having *y* goals

Expected Value (Mean): $E\{Y\} = \mu_Y = \sum_{y=0}^{7} yp(y) = 0(.0455) + ... + 7(.0227) = 2.7045$

Variance: $\sigma_Y^2 = \sigma^2\{Y\} = E\left\{(Y-\mu)^2\right\} = E\{Y^2\} - \mu^2 = \sum_{y=0}^{7} y^2 p(y) - \mu^2 = 9.9091 - 2.7045^2 = 2.5945$

Standard Deviation: $\sigma_Y = \sigma\{Y\} = +\sqrt{\sigma^2\{Y\}} = \sqrt{2.5945} = 1.6108$

<u>**Bivariate Random Variables**</u>

Joint Probability Function - Discrete Case (Generalizes to Densities in Continuous Case)

Random Variables (Outcomes observed on same unit) $\equiv Y, Z$ (k possibilities for Y, m for Z):

$g(Y_s, Z_t) = P(Y = Y_s \cap Z = Z_t) \quad s = 1, \ldots, k; t = 1, \ldots, m$ $\qquad$ Probability $Y = Y_s$ and $Z = Z_t$

Often written as $g(y, z)$ for specific outcomes y, z

Marginal Probability Function - Discrete Case (Generalizes to Densities in Continuous Case):

$f(Y_s) = \sum_{t=1}^{m} g(Y_s, Z_t)$ Probability $Y = Y_s$ $\qquad$ $h(Z_t) = \sum_{s=1}^{k} g(Y_s, Z_t)$ Probability $Z = Z_t$ $\qquad$ Often denoted $f(y)$, $\quad h(z)$

Continuous: Replace summations with integrals

Conditional Probability Function - Discrete Case (Generalizes to Densities in Continuous Case):

$f(Y_s \mid Z_t) = \dfrac{g(Y_s, Z_t)}{h(Z_t)} \quad h(Z_t) \neq 0; s = 1, \ldots, k$ $\qquad$ Probability $Y = Y_s$ given $Z = Z_t$ $\qquad$ Often denoted $f(y \mid z)$

$h(Z_t \mid Y_s) = \dfrac{g(Y_s, Z_t)}{f(Y_s)} \quad f(Y_s) \neq 0; t = 1, \ldots, m$ $\qquad$ Probability $Z = Z_t$ given $Y = Y_s$ $\qquad$ Often denoted $\mathrm{h}(z \mid y)$

<u>**Example – Goals by Half Y=Home Club Z=Away Club – Irish Premier League (2012)**</u>

H\A Freq	0	1	2	3	4	5	Total(Home)
0	105	67	20	8	0	0	200
1	75	41	18	1	0	0	135
2	26	17	1	0	1	0	45
3	6	3	3	0	0	0	12
4	1	1	0	0	0	0	2
5	2	0	0	0	0	0	2
Total(Away)	215	129	42	9	1	0	396
H\A Prob	0	1	2	3	4	5	Total(Home)
0	0.26515	0.16919	0.05051	0.02020	0.00000	0.00000	0.50505
1	0.18939	0.10354	0.04545	0.00253	0.00000	0.00000	0.34091
2	0.06566	0.04293	0.00253	0.00000	0.00253	0.00000	0.11364
3	0.01515	0.00758	0.00758	0.00000	0.00000	0.00000	0.03030
4	0.00253	0.00253	0.00000	0.00000	0.00000	0.00000	0.00505
5	0.00505	0.00000	0.00000	0.00000	0.00000	0.00000	0.00505
Total(Away)	0.54293	0.32576	0.10606	0.02273	0.00253	0.00000	1.00000

Home Team Distribution: f(y)

Away Team Distribution: g(z)

To obtain the conditional distribution of Away goals given a particular number of Home Goals, take the cell probabilities and divide by the total row probability. Similarly, for the conditional distribution of Home goals given Away goals, divide cell by column total.

Conditional Distribution of Home goals given Away Goals=0 ≡ f(y|z=0):

$$f(y=0|z=0)=\frac{0.26515}{0.54293}=0.48837 \quad f(y=1|z=0)=\frac{0.18939}{0.54293}=0.34884 \quad f(y=2|z=0)=\frac{0.06566}{0.54293}=0.12093$$
$$f(y=3|z=0)=\frac{0.01515}{0.54293}=0.02791 \quad f(y=4|z=0)=\frac{0.00253}{0.54293}=0.00465 \quad f(y=5|z=0)=\frac{0.00505}{0.54293}=0.00930$$
Note: $0.48837+0.34884+0.12093+0.02791+0.00465+0.00930=1$

Covariance, Correlation, and Independence

Covariance - Average of Product of Distances from Means

$$\sigma_{YZ}=\sigma\{Y,Z\}=E\{(Y-E\{Y\})(Z-E\{Z\})\}=E\{(Y-\mu_Y)(Z-\mu_Z)\}$$

Equivalently (for computing): $\sigma_{YZ}=\sigma\{Y,Z\}=E\{YZ\}-(E\{Y\})(E\{Z\})=E\{YZ\}-\mu_Y\mu_Z$

Note: Discrete: $E\{YZ\}=\sum_{s=1}^{k}\sum_{t=1}^{m}Y_sZ_tg(Y_s,Z_t)$ (Replace summations with integrals in continuous case)

a_1,c_1,a_2,c_2 are constants $\Rightarrow \sigma\{a_1+c_1Y,a_2+c_2Z\}=c_1c_2\sigma_{YZ}=c_1c_2\sigma\{Y,Z\}$

$\Rightarrow \sigma\{c_1Y,c_2Z\}=c_1c_2\sigma_{YZ}=c_1c_2\sigma\{Y,Z\} \quad \Rightarrow \sigma\{a_1+Y,a_2+Z\}=\sigma_{YZ}=\sigma\{Y,Z\}$

Correlation: Covariance scaled to lie between -1 and +1 for measure of association strength

Standardized Random Variables (Scaled to have mean=0, variance=1) $Y'=\frac{Y-E\{Y\}}{\sigma\{Y\}}=\frac{Y-\mu_Y}{\sigma_Y}$

$$\rho_{YZ}=\rho\{Y,Z\}=\sigma\{Y',Z'\}=\frac{\sigma\{Y,Z\}}{\sigma\{Y\}\sigma\{Z\}} \qquad -1\le\rho\{Y,Z\}\le 1$$

$\sigma\{Y,Z\}=\rho\{Y,Z\}=0\Rightarrow Y,Z$ are uncorrelated (not necessarily independent)

Independent Random Variables

Y,Z are independent if and only if $g(Y_s,Z_t)=f(Y_s)h(Z_t) \quad s=1,\ldots,k;\ t=1,\ldots,m$

If Y,Z are jointly normally distributed and $\sigma\{Y,Z\}=0$ then Y,Z are independent

Average Home Goals per Half: $\mu_Y=0(0.50505)+\ldots+5(.00505)=0.70455$

Average Away Goals per Half: $\mu_Z=0(0.54293)+\ldots+5(.00000)=0.61616$

$E\{Y^2\}=0^2(0.50505)+\ldots+5^2(.00505)=1.27525$

$E\{Z^2\}=0^2(0.54293)+\ldots+5^2(.00000)=0.99495$

$E\{YZ\}=0(0)(0.26515)+0(1)(0.16919)+\ldots+5(5)(0.00000)=0.39647$

$\sigma_Y^2=1.27525-0.70455^2=0.77887 \qquad \sigma_Y=\sqrt{0.77887}=0.88254$

$\sigma_Z^2=0.99495-0.61616^2=0.61529 \qquad \sigma_Z=\sqrt{0.61529}=0.78441$

$\sigma_{YZ}=\sigma\{Y,Z\}=E\{YZ\}-\mu_Y\mu_Z=0.39647-0.70455(0.61616)=-0.03765$

$$\rho_{YZ}=\frac{\sigma_{YZ}}{\sigma_Y\sigma_Z}=\frac{-0.03765}{0.88254(0.78441)}=-0.05439$$

To see that Home and Away Goals are NOT independent (besides simply observing the correlation is not zero), you can check whether the joint probabilities in the cells of the joint distribution are all equal to the product of their row and column totals (product of the marginal probabilities).

For the case where both Home and Away goals are 0:

$$g(y=0,z=0)=0.26515 \quad f(y=0)=0.50505 \quad h\left(z=0\right)=0.54293$$
$$0.26515 \neq 0.50505(0.54293)=0.27421$$

Linear Functions of Random Variables

$$U=\sum_{i=1}^{n} a_iY_i \quad \{a_i\}\equiv \text{constants} \quad \{Y_i\}\equiv \text{random variables}$$
$$E\{Y_i\}=\mu_i \quad \sigma^2\{Y_i\}=\sigma_i^2 \quad \sigma\{Y_i,Y_j\}=\sigma_{ij}$$
$$\Rightarrow \quad E\{U\}=E\left\{\sum_{i=1}^{n} a_iY_i\right\}=\sum_{i=1}^{n} a_iE\{Y_i\}=\sum_{i=1}^{n} a_i\mu_i$$
$$\Rightarrow \quad \sigma^2\{U\}=\sigma^2\left\{\sum_{i=1}^{n} a_iY_i\right\}=\sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\sigma_{ij}=\sum_{i=1}^{n} a_i^2\sigma_i^2+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_ia_j\sigma_{ij}$$
$$n=2\Rightarrow E\{a_1Y_1+a_2Y_2\}=a_1E\{Y_1\}+a_2E\{Y_2\}=a_1\mu_1+a_2\mu_2$$
$$\sigma^2\{a_1Y_1+a_2Y_2\}=a_1^2\sigma^2\{Y_1\}+a_2^2\sigma^2\{Y_2\}+2a_1a_2\sigma\{Y_1,Y_2\}=a_1^2\sigma_1^2+a_2^2\sigma_2^2+2a_1a_2\sigma_{12}$$

Total Goals, Difference (Home – Away), and Average Goals by Half Y_1 = Home Y_2 = Away:

$$\mu_1=\mu_Y=0.70455 \quad \mu_2=\mu_Z=0.61616 \quad \sigma_1^2=\sigma_Y^2=0.77887 \quad \sigma_2^2=\sigma_Z^2=0.61529 \quad \sigma_{12}=\sigma_{YZ}=-0.03765$$
$$\text{Total Goals: } U_1=Y_1+Y_2 \quad (a_1=1,a_2=1)$$
$$\text{Difference in Goals: } U_2=Y_1-Y_2 \quad (a_1=1,a_2=-1)$$
$$\text{Average Goals: } U_3=\frac{Y_1+Y_2}{2} \quad \left(a_1=\frac{1}{2},a_2=\frac{1}{2}\right)$$
$$\mu_{U_1}=1\mu_1+1\mu_2=1(0.70455)+1(0.61616)=1.32071$$
$$\sigma_{U_1}^2=1^2\sigma_1^2+1^2\sigma_2^2+2(1)(1)\sigma_{12}=1(0.77887)+1(0.61529)+2(-0.03765)=1.31886$$
$$\mu_{U_2}=1\mu_1+(-1)\mu_2=1(0.70455)-1(0.61616)=0.08838$$
$$\sigma_{U_2}^2=1^2\sigma_1^2+(-1)^2\sigma_2^2+2(1)(-1)\sigma_{12}=1(0.77887)+1(0.61529)-2(-0.03765)=1.469461$$
$$\mu_{U_3}=\frac{1}{2}\mu_1+\frac{1}{2}\mu_2=\frac{1}{2}(0.70455)+\frac{1}{2}(0.61616)=0.66035$$
$$\sigma_{U_3}^2=\left(\frac{1}{2}\right)^2\sigma_1^2+\left(\frac{1}{2}\right)^2\sigma_2^2+2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\sigma_{12}=\frac{1}{4}(0.77887)+\frac{1}{4}(0.61529)+\frac{1}{2}(-0.03765)=0.32972$$

Linear Functions of INDEPENDENT Random Variables

$$Y_1,\ldots,Y_n \equiv \text{independent} \;\Rightarrow\; \sigma^2\left\{U\right\} = \sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

Special Cases $\left(Y_1, Y_2 \text{ independent}\right)$:

$$U_1 = Y_1 + Y_2 \qquad \sigma^2\left\{U_1\right\} = \sigma^2\left\{Y_1 + Y_2\right\} = (1)^2\sigma_1^2 + (1)^2\sigma_2^2 = \sigma_1^2 + \sigma_2^2$$

$$U_2 = Y_1 - Y_2 \qquad \sigma^2\left\{U_2\right\} = \sigma^2\left\{Y_1 - Y_2\right\} = (1)^2\sigma_1^2 + (-1)^2\sigma_2^2 = \sigma_1^2 + \sigma_2^2$$

$$Y_1,\ldots,Y_n \equiv \text{independent} \;\Rightarrow\; \sigma^2\left\{\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i\right\} = \sum_{i=1}^{n} a_i c_i \sigma_i^2$$

Special Case $\left(Y_1, Y_2 \text{ independent}\right)$:

$$\sigma\left\{U_1, U_2\right\} = \sigma\left\{Y_1 + Y_2, Y_1 - Y_2\right\} = (1)(1)\sigma_1^2 + (1)(-1)\sigma_2^2 = \sigma_1^2 - \sigma_2^2$$

Note: These do not apply for the soccer data, but are used repeatedly to obtain properties of estimators in linear models.

Central Limit Theorem

When random samples of size *n* are selected from any population with mean *m* and finite variance s^2, the sampling distribution of the sample mean will be approximately normally distributed for large *n*:

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \sum_{i=1}^{n}\left(\frac{1}{n}\right) Y_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\text{approximately, for large } n$$

Z-table (and software packages) can be used to approximate probabilities of ranges of values for sample means, as well as percentiles of their sampling distribution

Probability Distributions Widely Used in Linear Models

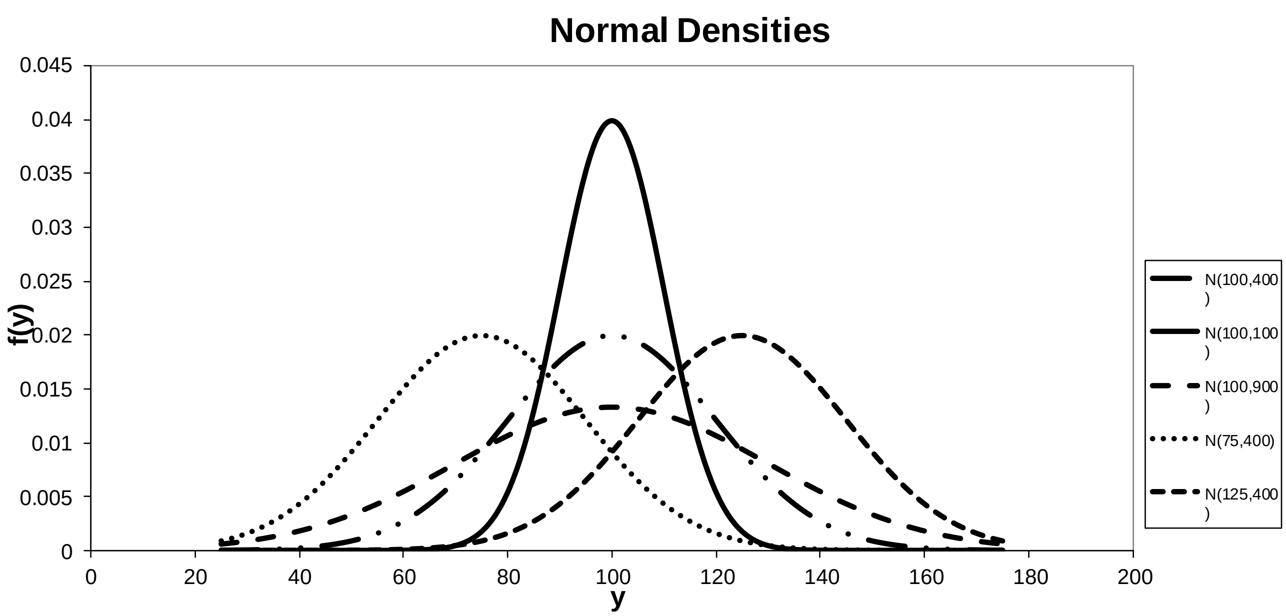
Normal (Gaussian) Distribution

- **Bell-shaped distribution with tendency for individuals to clump around the group median/mean**
- **Used to model many biological phenomena**
- **Many estimators have approximate normal sampling distributions (see Central Limit Theorem)**
- **Notation: $Y \sim N(\mu,\sigma^2)$ where μ is mean and σ^2 is variance**

$$\boxed{f(y)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{(y-\mu)^2}{\sigma^2}\right)\right] \quad -\infty<y<\infty, -\infty<\mu<\infty, \sigma>0}$$

Probabilities can be obtained from software packages (e.g. EXCEL, R, SPSS, SAS, STATA). Tables can be used to obtain probabilities once values have been standardized to have mean 0, and standard deviation 1.

$$Y \sim N\left(\mu_Y,\sigma_Y^2\right) \quad \Rightarrow \quad Z=\frac{Y-\mu_Y}{\sigma_Y} \sim N\left(\mu_Z=0,\sigma_Z^2=1\right)$$



EXCEL Commands for Probabilities and Quantiles (Default are lower tail areas):

- **Lower tail (cumulative) probabilities: =norm.dist(y,mu,sigma,True)**
- **Upper tail probabilities: =1 - norm.dist(y,mu,sigma,True)**
- **p^{th} quantile: =norm.inv(p,mu,sigma) 0<p<1**

Second Decimal Place of Z

Integer and first decimal place

F(z)	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table gives $F(z) = P(Z \leq z)$ for a wide range of z-values (0 to 3.09 by 0.01)

Notes:

- $P(Z \geq z) = 1-F(z)$
- $P(Z \leq -z) = 1-F(z)$
- $P(Z \geq -z) = F(z)$

R Program to Obtain Probabilities, Percentiles, Density Functions, and Random Sampling

```
# Obtain P(Y<=80|N(mu=100,sigma=20))
# pnorm gives lower tail probabilities (cdf) for a normal distribution
pnorm(80,mean=100,sd=20)

# Obtain P(Y>=80|N(mu=100,sigma=20))
# lower=FALSE option gives upper tail probabilities
pnorm(80,mean=100,sd=20,lower=FALSE)

# Obtain the 10th percentile of a Normal Density with mu=100, sigma=20
qnorm(0.10, mean=100, sd=20)

# Obtain a plot of a Normal Density with mu=100, sigma=20
# dnorm gives the density function for a normal distribution at point(s) y
# type="l" in plot function joins the points on the density function with a line
# The polygon command fills in the area below y=80 in green
 y <- seq(40,160,0.01)
fy <- dnorm(y,mean=100,sd=20)

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\norm_dist1.png")

plot(y,fy,type="l",
main=expression(paste("Normal(",mu,"=100,",sigma,"=20)")))
polygon(c(y[y<=80],80),c(fy[y<=80],fy[y==40]),col="green")

dev.off()   # Close the .png file

# Obtain a random sample of 1000 items from N(mu=100,sigma=20)
# rnorm gives a random sample of size given by the first argument
# Obtain sample mean, median, variance, standard deviation

set.seed(54321)       # Set the seed for random number generator for reproducing data
y.samp <- rnorm(1000,mean=100,sd=20)
mean(y.samp)
median(y.samp)
var(y.samp)
sd(y.samp)

# Plot a histogram of the sample values (Default bin size)
hist(y.samp, main = expression(paste("Sampled values, ", mu, "=100, ", sigma,
   "=20")))

# Allow for more bins

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\norm_dist2.png")

hist(y.samp, breaks=23,
main = expression(paste("Sampled values, ", mu, "=100, ", sigma,
   "=20")))

# Add normal density (scaled up by (n=1000 x binwidth=5), since a freq histogram)
# Makes use of y and fy defined above

lines(y,1000*5*fy)

dev.off()   # Close the .png file
```

Numeric Output from R Program

```
>
> pnorm(80,mean=100,sd=20)
[1] 0.1586553
>
> pnorm(80,mean=100,sd=20,lower=FALSE)
[1] 0.8413447
>
> qnorm(0.10, mean=100, sd=20)
[1] 74.36897

> mean(y.samp)
[1] 98.80391
> median(y.samp)
[1] 98.95658
> var(y.samp)
[1] 407.2772
> sd(y.samp)
[1] 20.18111
```

Note that the first 3 values are easily computed using the z-table. The last 4 values would take lots of calculations based on a sample of 1000 observations.

$$Y \sim N\left(\mu = 100, \sigma^2 = 20^2 = 400\right)$$

$$P\left(Y \leq 80\right) = P\left(Z = \frac{Y-\mu}{\sigma} \leq \frac{80-100}{20} = -1\right) = 1 - P\left(Z \geq 1\right) = 1 - .8413 = .1587$$

$$P\left(Y \geq 80\right) = P\left(Z \geq -1\right) = P\left(Z \leq 1\right) = .8413$$

$10th$-Percentile: From z-table: $P\left(Z \leq -1.28\right) = 1 - P\left(Z \leq 1.28\right) = 1 - .8997 = .1003 \approx .10$

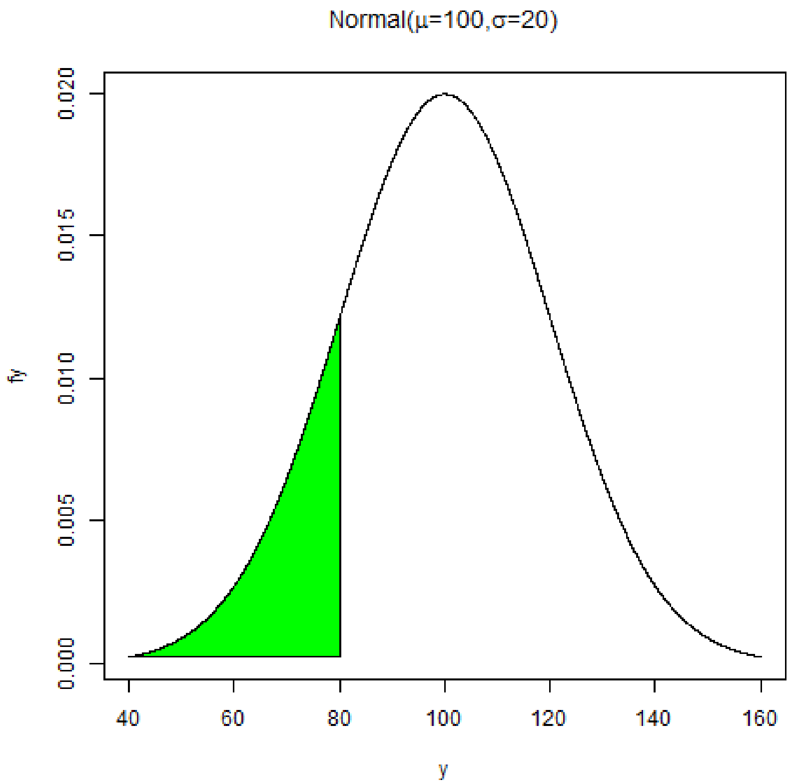
$$.10 \approx P\left(Z \leq -1.28\right) = P\left(Z = \frac{Y-\mu}{\sigma} \leq -1.28\right) = P\left(Y \leq -1.28\sigma + \mu\right) = P\left(Y \leq -1.28(20) + 100 = 74.4\right)$$

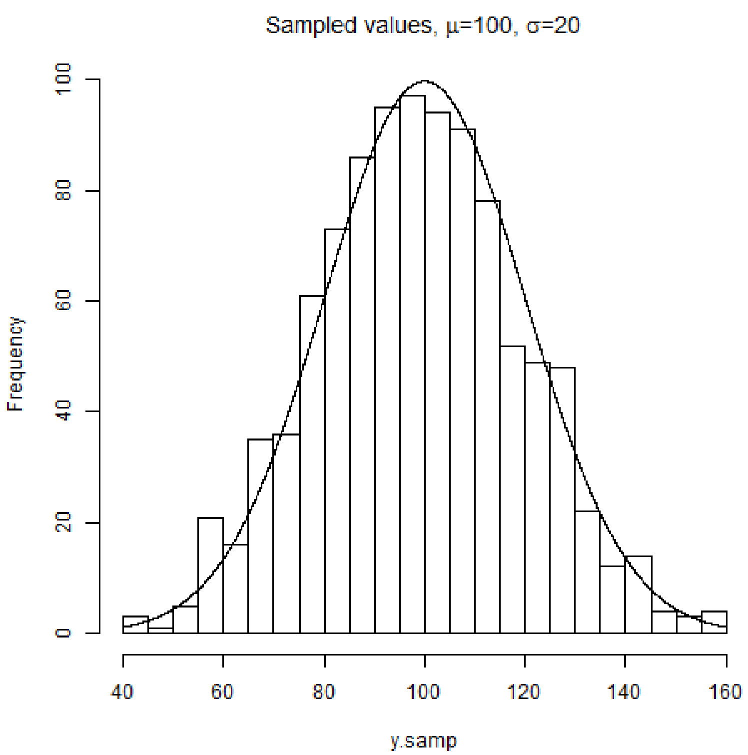
Cell	Result
A1	0.158655
A2	0.841345
A3	74.36897

EXCEL Output:

- **Cell A1: =NORM.DIST(80,100,20,TRUE)**
- **Cell A2: =1-NORM.DIST(80,100,20,TRUE)**
- **Cell A3: =NORM.INV(0.1,100,20)**

Graphics Output from R Program





Chi-Square Distribution

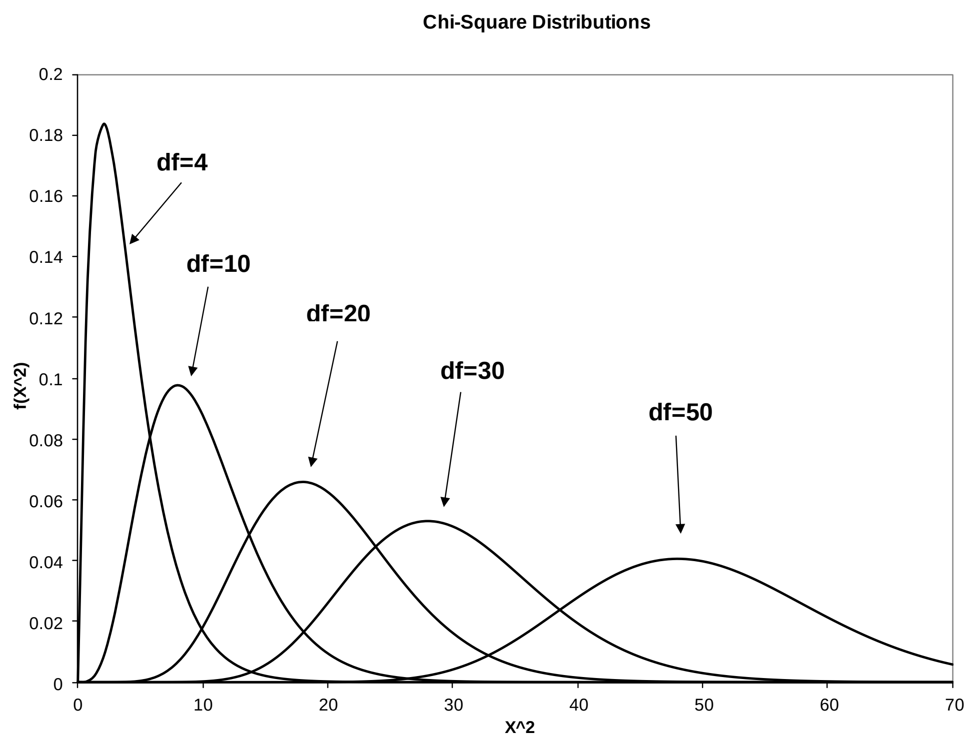
- **Indexed by "degrees of freedom (ν)" $X \sim \chi_\nu^2$**
- **$Z \sim N(0,1) \Rightarrow Z^2 \sim \chi_1^2$**
- **Assuming Independence:**

$$X_1,\ldots,X_n \sim \chi^2_{\nu_i} \quad i=1,\ldots,n \quad \Rightarrow \quad \sum_{i=1}^{n} X_i \sim \chi^2_{\sum \nu_i}$$

Density Function:

$$f\left(x\right)=\frac{1}{\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}}x^{(\nu/2)-1}e^{-x/2} \quad x>0,\nu>0$$

Probabilities can be obtained from software packages (e.g. EXCEL, R, SPSS, SAS, STATA). Tables can be used to obtain certain critical values for given upper and lower tail areas.



EXCEL Commands for Probabilities and Quantiles (Default are upper tail areas):

- **Lower tail (cumulative) probabilities: =1-chidist(y,df)**
- **Upper tail probabilities: = chidist(y,df)**
- **p^{th} quantile: =chiinv(1-p,df) 0<p<1**

Critical Values for Chi-Square Distributions (Mean=ν, Variance=2ν)

df\F(x)	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

R Program to Obtain Probabilities, Percentiles, Density Functions, and Random Sampling

```
# Obtain P(Y<=5|X2(df=10))
# pchisq gives lower tail probabilities (cdf) for a chi-square distribution
pchisq(5,df=10)


# Obtain P(Y>=5|X2(df=10))
# lower=FALSE option gives upper tail probabilities
pchisq(5,df=10,lower=FALSE)

# Obtain the 95th percentile of a Chi-square Density with df=10
qchisq(0.95,df=10)

# Obtain a plot of a Chi-square Density with df=10
# dchisq gives the density function for a chi-square distribution at point(s) y
# type="l" in plot function joins the points on the density function with a line
# The polygon command fills in the area below y<5 in green

y <- seq(0,30,0.01)
fy <- dchisq(y,df=10)

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\chisq_dist1.png")

plot(y,fy,type="l",
main=expression(paste(chi^2,"(df=10)")))
polygon(c(y[y<=5],5),c(fy[y<=5],fy[y==0]),col="blue")

dev.off()   # Close the .png file


# Obtain a random sample of 1000 items from Chi-square(df=10)
# rchisq gives a random sample of size given by the first argument
# Obtain sample mean, median, variance, standard deviation

set.seed(54321)       # Set the seed for random number generator for reproducing data
y.samp <- rchisq(1000,df=10)
mean(y.samp)
median(y.samp)
var(y.samp)
sd(y.samp)

# Plot a histogram of the sample values (Default bin size)
hist(y.samp, main = expression(paste("Sampled values, ", chi^2, "(df=10)")))

# Allow for more bins

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\chisq_dist2.png")

hist(y.samp[y.samp<=30], breaks=29,
main = expression(paste("Sampled values, ", chi^2, "(df=10)")))

# Add chi-square density (scaled up by (n=1000 x binwidth=1), since a freq histogram)
# Makes use of y and fy defined above

lines(y,1000*1*fy)

dev.off()   # Close the .png file
```

Numeric Output from R Program

```
>
> pchisq(5,df=10)
[1] 0.108822
>
> pchisq(5,df=10,lower=FALSE)
[1] 0.891178
>
> qchisq(0.95,df=10)
[1] 18.30704

> mean(y.samp)
[1] 9.834778
> median(y.samp)
[1] 9.060967
> var(y.samp)
[1] 21.78964
> sd(y.samp)
[1] 4.667937
```

Note that for the chi-square distribution, the mean is the degrees of freedom (ν) and the variance is 2ν. The sample mean and variance are close to 10 and 20. As the sample size gets larger, they will tend to get closer. Also notice that the median is lower than the mean (right-skewed distribution).

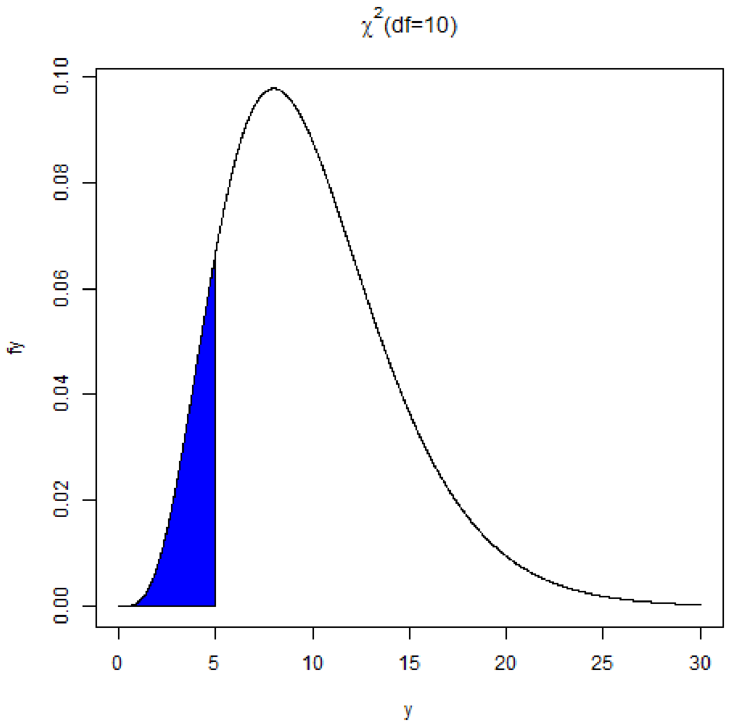
Confirm that the 95th-percentile is consistent with the table value.

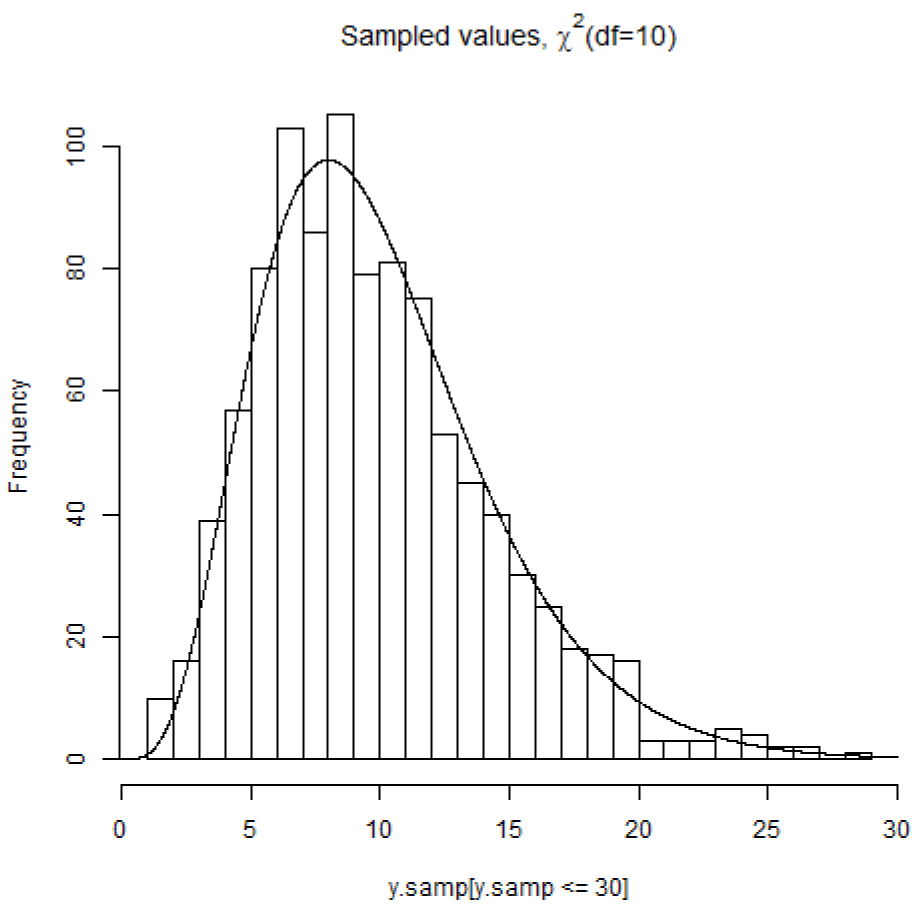
Cell	Result
A1	0.108822
A2	0.891178
A3	18.30704

EXCEL Output:

- **Cell A1: =1-CHIDIST(5,10)**
- **Cell A2: =CHIDIST(5,10)**
- **Cell A3: =CHIINV(0.05,10)**

Graphics Output from R Program



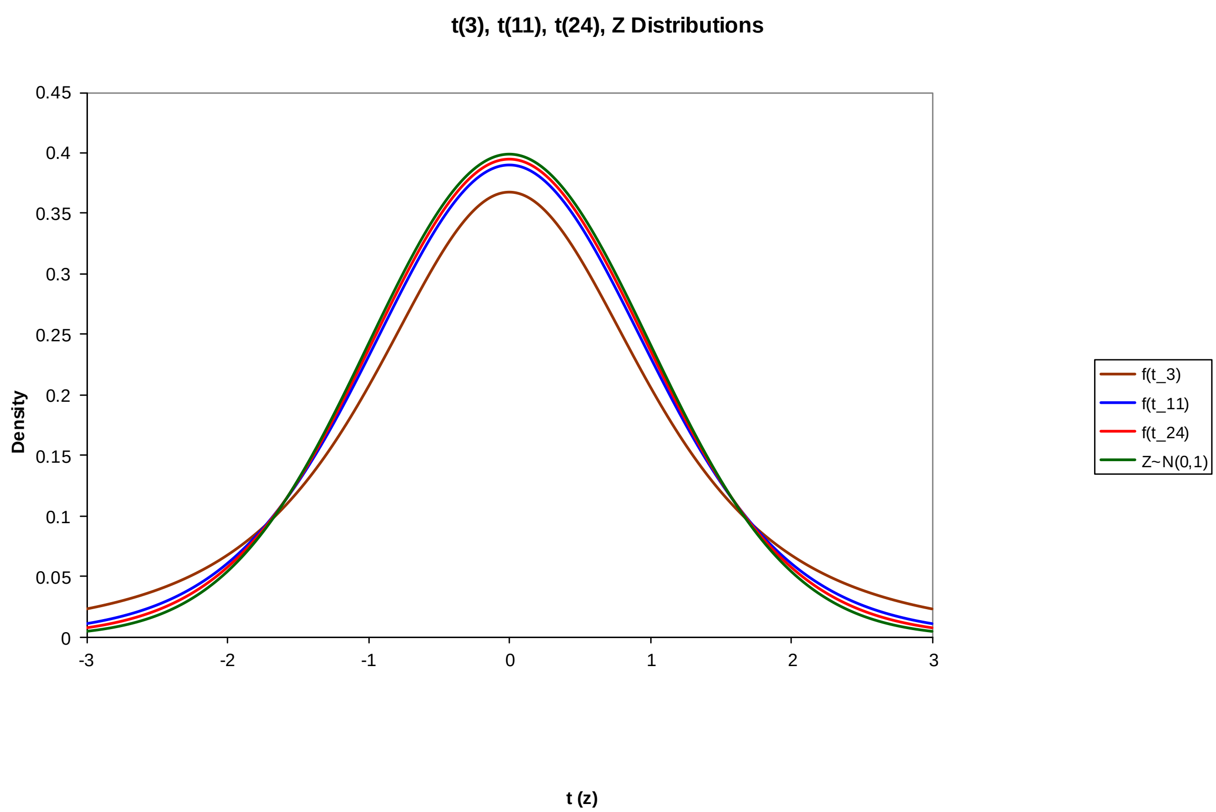


Student's t-Distribution

- **Indexed by "degrees of freedom (ν)" $X \sim t_\nu$**
- **$Z \sim N(0,1)$, $X \sim \chi_n^2$**
- **Assuming Independence of Z and X:**

$$\boxed{T = \frac{Z}{\sqrt{X/\nu}} \sim t(\nu)}$$

Probabilities can be obtained from software packages (e.g. EXCEL, R, SPSS, SAS, STATA). Tables can be used to obtain certain critical values for given upper tail areas (distribution is symmetric around 0, as N(0,1) is.



EXCEL Commands for Probabilities and Quantiles (Default are lower tail areas):

- **Lower tail (cumulative) probabilities: =t.dist(y,df,TRUE)**
- **Upper tail probabilities: =1- t.dist(y,df,TRUE)**
- **p^{th} quantile: =t.inv(p,df) 0<p<1**

Critical Values for Student's t-Distributions (Mean=0, Variance = $\nu/(\nu-2)$, $\nu>2$)

df\F(t)	0.9	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626

R Program to Obtain Probabilities, Percentiles, Density Functions, and Random Sampling

```
# Obtain P(Y<=1|t(df=8))
# pt gives lower tail probabilities (cdf) for a t distribution
pt(1,df=8)


# Obtain P(Y>=1|t(df=8))
# lower=FALSE option gives upper tail probabilities
pt(1,df=8,lower=FALSE)

# Obtain the 90th percentile of a t Density with df=8
qt(0.90,df=8)

# Obtain a plot of a t Density with df=8
# dt gives the density function for a tdistribution at point(s) y
# type="l" in plot function joins the points on the density function with a line
# The polygon command fills in the area below y<1 in red
 y <- seq(-4,4,0.01)
fy <- dt(y,df=8)

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\t_dist1.png")
plot(y,fy,type="l",
main="t(df=8)")
polygon(c(y[y<=1],1),c(fy[y<=1],fy[y==-4]),col="red")
dev.off()   # Close the .png file


# Obtain a random sample of 1000 items from t(df=8)
# rt gives a random sample of size given by the first argument
# Obtain sample mean, median, variance, standard deviation

set.seed(54321)       # Set the seed for random number generator for reproducing data
y.samp <- rt(1000,df=8)
mean(y.samp)
median(y.samp)
var(y.samp)
sd(y.samp)

# Plot a histogram of the sample values (Default bin size)
hist(y.samp, main ="Sampled values, t(df=8)")

# Allow for more bins

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\t_dist2.png")

hist(y.samp[abs(y.samp)<=4], breaks=31,
main ="Sampled values, t(df=8)")


# Add t density (scaled up by (n=1000 x binwidth=0.25), since a freq histogram)
# Makes use of y and fy defined above

lines(y,1000*0.25*fy)

dev.off()   # Close the .png file
```

Numeric Output from R Program

```
> pt(1,df=8)
[1] 0.8267032
>
> pt(1,df=8,lower=FALSE)
[1] 0.1732968
>
> qt(0.90,df=8)
[1] 1.396815

> mean(y.samp)
[1] -0.03754771
> median(y.samp)
[1] 0.0007432709
> var(y.samp)
[1] 1.43555
> sd(y.samp)
[1] 1.198145
```

Note that for the t distribution, the mean is 0, and the variance is $\nu/(\nu-2)$. The sample mean and variance are close to 0 and 8/6=1.333. As the sample size gets larger, they will tend to get closer.

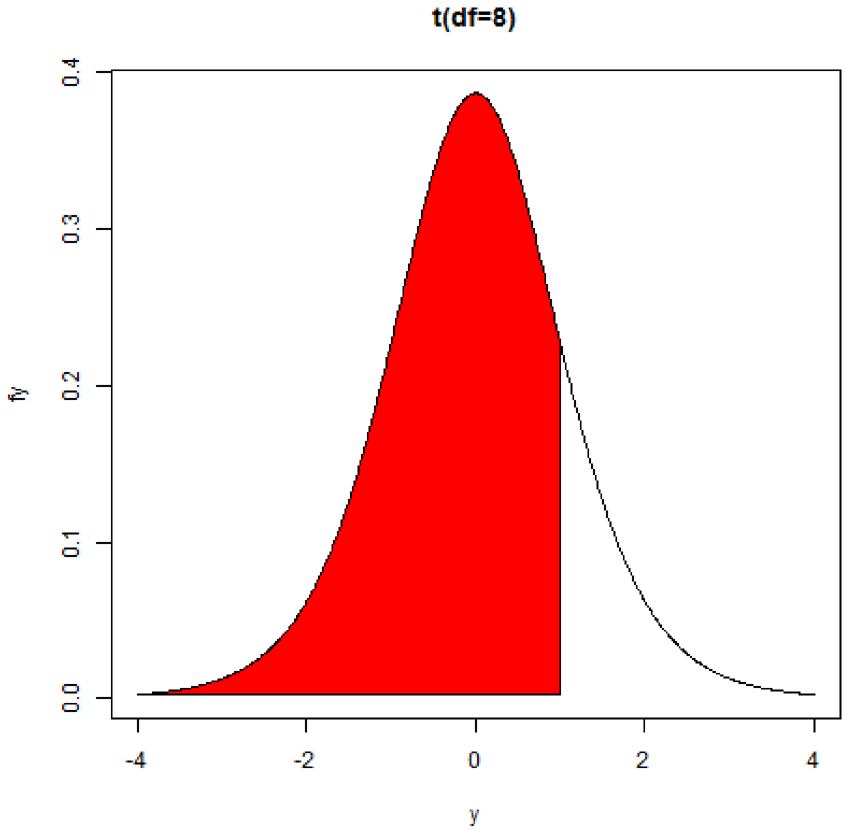
Confirm that the 90^{th}-percentile is consistent with the table value.

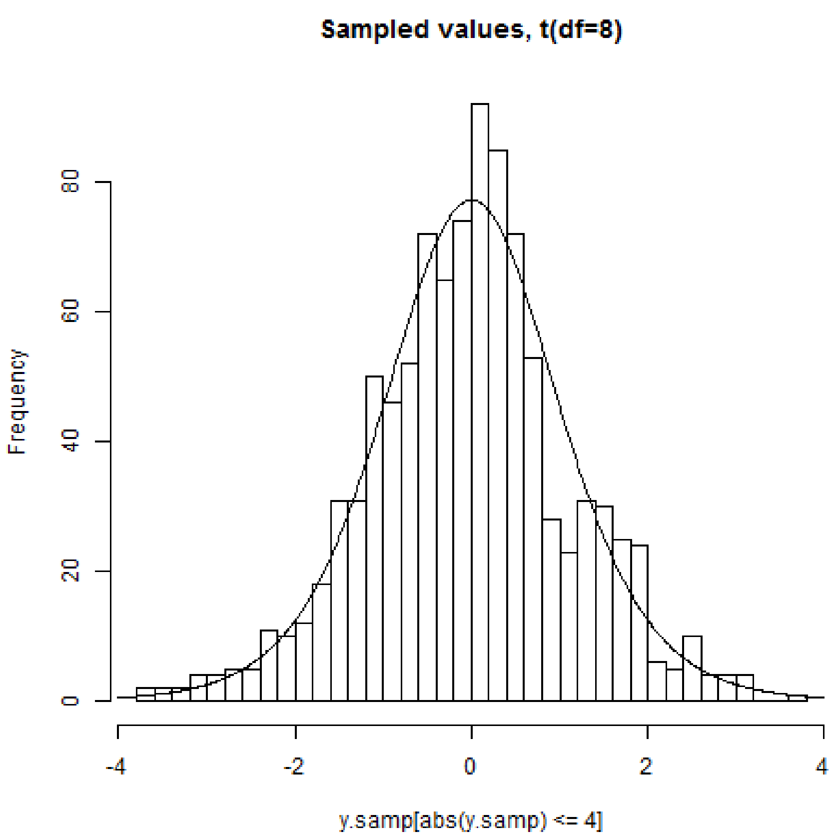
Cell	Result
A1	0.826703
A2	0.173297
A3	1.396815

EXCEL Output:

- **Cell A1: =T.DIST(1,8,TRUE)**
- **Cell A2: =1-T.DIST(1,8,TRUE)**
- **Cell A3: =T.INV(0.9,8)**

Graphics Output from R Program



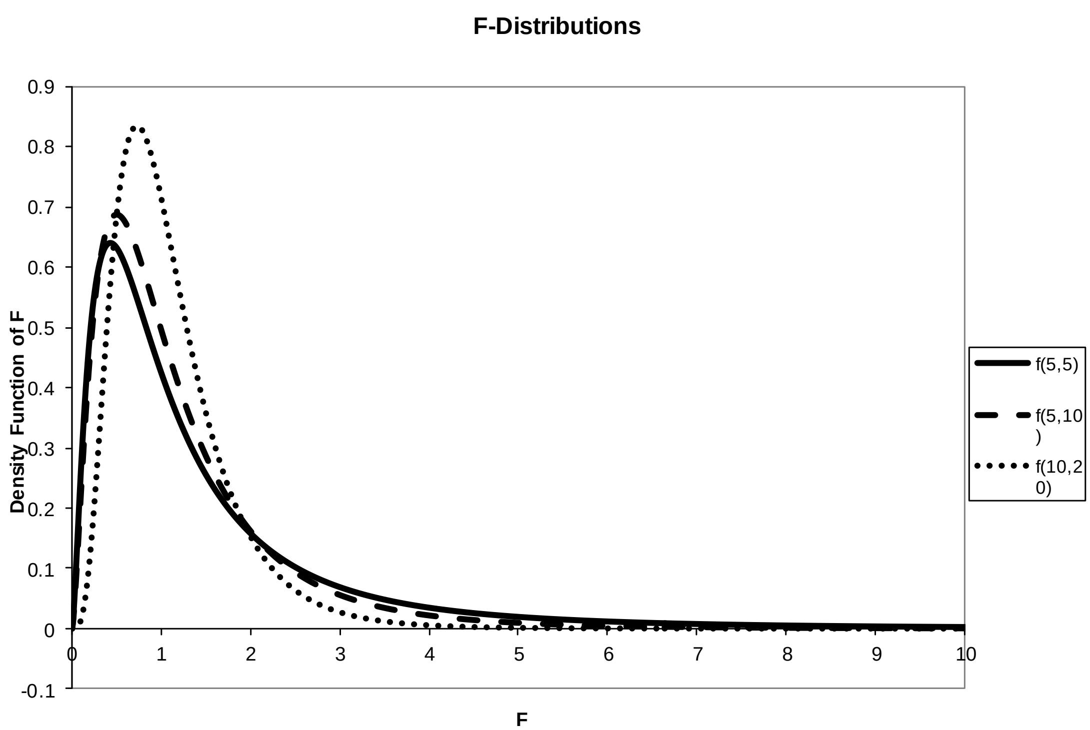


F-Distribution

- **Indexed by 2 "degrees of freedom (ν_1,ν_2)" W~$F_{\nu 1,\nu 2}$**
- **$X_1 \sim \chi_{\nu 1}^2$, $X_2 \sim \chi_{\nu 2}^2$**
- **Assuming Independence of X_1 and X_2:**

$$\boxed{W = \frac{X_1/\nu_1}{X_2/\nu_2} \sim F\left(\nu_1, \nu_2\right)}$$

Probabilities can be obtained from software packages (e.g. EXCEL, R, SPSS, SAS, STATA). Tables can be used to obtain certain critical values for given upper tail areas. Lower tails are obtained by taking the reciprocal of the upper tail with the degrees of freedom reversed.



EXCEL Commands for Probabilities and Quantiles (Default are upper tail areas):

- **Lower tail (cumulative) probabilities: =1-fdist(y,df1,df2)**
- **Upper tail probabilities: = fdist(y,df1,df2)**
- **p^{th} quantile: =finv(1-p,df1,df2) 0<p<1**

Critical Values for F-distributions   P(F ≤ Table Value) = 0.95

df2\df1	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93

R Program to Obtain Probabilities, Percentiles, Density Functions, and Random Sampling

```
# Obtain P(Y<=2.5|F(df1=10,df2=8))
# pf gives lower tail probabilities (cdf) for a F distribution
pf(2.5,df1=10,df2=8)


# Obtain P(Y>=2.5|F(df1=10,df2=8)))
# lower=FALSE option gives upper tail probabilities
pf(2.5,df1=10,df2=8,lower=FALSE)

# Obtain the 5th and 95th percentiles of a F Density with df1=10,df2=8
qf(0.05,df1=10,df2=8)
qf(0.95,df1=10,df2=8)

# Obtain a plot of a F Density with df1=10, df2=8
# df gives the density function for a F distribution at point(s) y
# type="l" in plot function joins the points on the density function with a line
# The polygon command fills in the area below y<2.5 in purple
 y <- seq(0,10,0.01)
fy <- df(y,df1=10,df2=8)

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\f_dist1.png")
plot(y,fy,type="l",
main="F(df1=10,df2=8)")
polygon(c(y[y<=2.5],2.5),c(fy[y<=2.5],fy[y==0]),col="purple")
dev.off()   # Close the .png file


# Obtain a random sample of 1000 items from F(df1=10,df2=8)
# rf gives a random sample of size given by the first argument
# Obtain sample mean, median, variance, standard deviation

set.seed(54321)       # Set the seed for random number generator for reproducing data
y.samp <- rf(1000,df1=10,df2=8)
mean(y.samp)
median(y.samp)
var(y.samp)
sd(y.samp)

# Plot a histogram of the sample values (Default bin size)

hist(y.samp, main ="Sampled values, F(df1=10,df2=8)")

# Allow for more bins

# Output graph to a .png file in the following directory/file)
png("E:\\blue_drive\\Rmisc\\graphs\\f_dist2.png")

hist(y.samp[y.samp<=10], breaks=19, ylim=c(0,400),
main ="Sampled values, F(df1=10,df2=8)")


# Add chi-square density (scaled up by (n=1000 x binwidth=0.5), since a freq histogram)
# Makes use of y and fy defined above

lines(y,1000*0.5*fy)

dev.off()   # Close the .png file
```

Numeric Output from R Program

```
> pf(2.5,df1=10,df2=8)
[1] 0.8964058
>
> pf(2.5,df1=10,df2=8,lower=FALSE)
[1] 0.1035942
>
> qf(0.05,df1=10,df2=8)
[1] 0.325557
> qf(0.95,df1=10,df2=8)
[1] 3.347163

> mean(y.samp)
[1] 1.369505
> median(y.samp)
[1] 1.059021
> var(y.samp)
[1] 1.50341
> sd(y.samp)
[1] 1.226136
```

Note that for the F distribution, the mean and variance formulas are given below.

Mean: $\dfrac{\nu_2}{\nu_2 - 2} \quad (\nu_2 > 2)$ Variance: $\dfrac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad (\nu_2 > 4)$

For this case, the mean is 8/6 = 1.333 and the variance is 2048/1440 = 1.422. Again the sample mean and variance would tend to be closer to the theoretical values as the sample size increases.

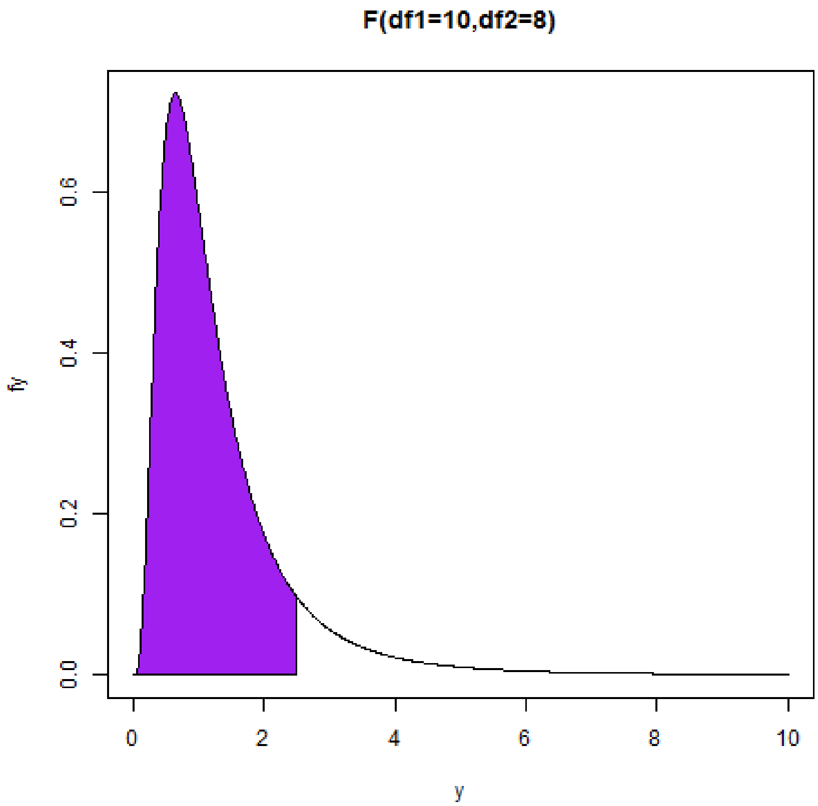
Confirm the 5th and 95th percentiles based on the F-table. Again note that the lower percentile can be obtained by taking the reciprocal of the upper percentile with the degrees of freedom reversed.

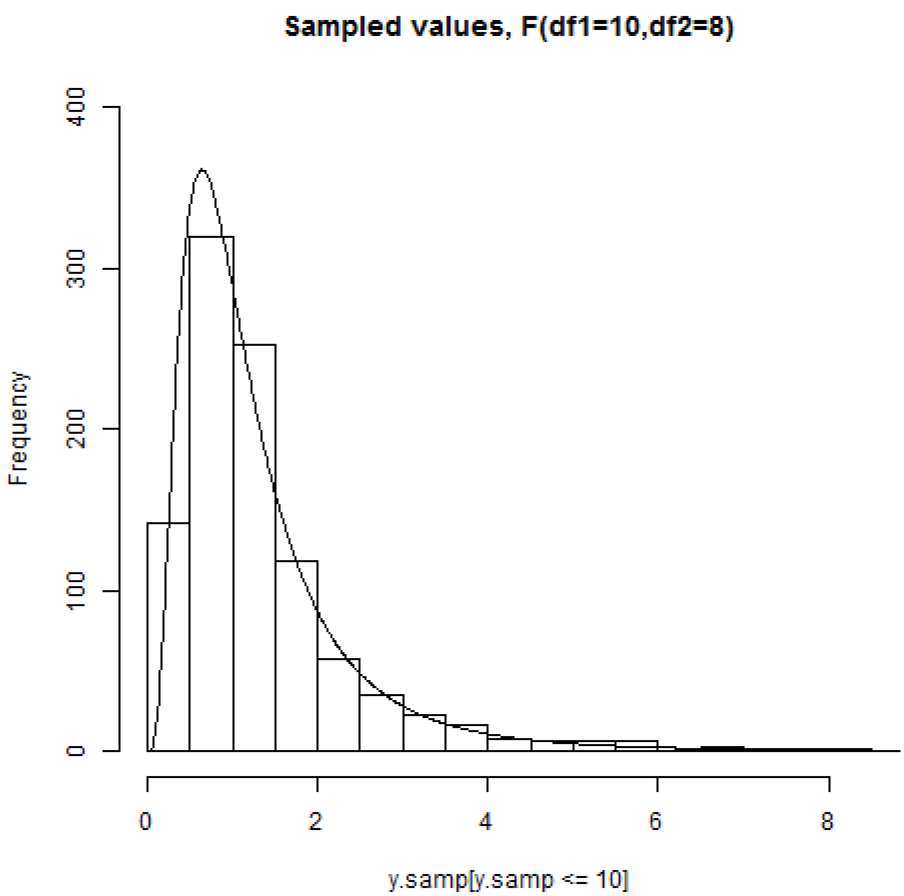
Cell	Result
A1	0.896406
A2	0.103594
A3	0.325557
A4	3.347163

EXCEL Output:

- **Cell A1: =1-FDIST(2.5,10,8)**
- **Cell A2: =FDIST(2.5,10,8)**
- **Cell A3: =FINV(0.95,10,8)**
- **Cell A4: =FINV(0.05,10,8)**

Graphics Output from R Program





Statistical Estimation: Properties

Properties of Estimators:

Parameter: θ Estimator: $\hat{\theta} \equiv$ function of $Y_1,...,Y_n$

1) Unbiased: $E\left\{\hat{\theta}\right\} = \theta$

2) Consistent: $\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \geq \varepsilon\right) = 0$ for any $\varepsilon > 0$

3) Sufficient if conditional joint probability of sample, given $\hat{\theta}$ does not depend on θ

4) Minimum Variance: $\sigma^2\left\{\hat{\theta}\right\} \leq \sigma^2\left\{\hat{\theta}^*\right\}$ for all $\hat{\theta}^*$

Note: If an estimator is unbiased (easy to show) and its variance goes to zero as its sample size gets infinitely large (easy to show), it is consistent. It is tougher to show that it is Minimum Variance, but general results have been obtained in many standard cases.

Statistical Estimation: Methods

Maximum Likelihood (ML) Estimators:

$Y \sim f(Y;\theta) \equiv$ Probability function for Y that depends on parameter θ

Random Sample (independent) $Y_1,...,Y_n$ with joint probability function:

$$g(Y_1,...,Y_n) = \prod_{i=1}^{n} f(Y;\theta)$$

When viewed as function of θ, given the observed data (sample):

Likelihood function: $L(\theta) = \prod_{i=1}^{n} f(Y;\theta)$ Goal: maximize $L(\theta)$ with respect to θ.

Under general conditions, ML estimators are consistent and sufficient

Least Squares (LS) Estimators

$Y_i = f_i(\theta) + \varepsilon_i$

where $f_i(\theta)$ is a known function of the parameter θ and ε_i are random variables, usually with $E\{\varepsilon_i\}$ =0

Sum of Squares: $Q = \sum_{i=1}^{n}\left[Y_i - f_i(\theta)\right]^2$ Goal: minimize Q with respect to θ.

In many settings, LS estimators are unbiased and consistent.

One-Sample Confidence Interval for μ

- **Simple Random Sample (SRS) from a population with mean μ is obtained.**
- **Sample mean, sample standard deviation are obtained**
- **Degrees of freedom are df= n-1, and confidence level (1-α) are selected**
- **Level (1-α) confidence interval of form:**

$$\overline{Y} \pm t\left(1-\alpha/2; n-1\right) s\left\{\overline{Y}\right\} \qquad s\left\{\overline{Y}\right\} = \frac{s}{\sqrt{n}} \qquad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad s^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-1}$$

Procedure is theoretically derived based on normally distributed data, but has been found to work well regardless for moderate to large n

Example: Mercury Levels Albacore Fish in the Eastern Mediterranean

Sample: n = 34 albacore fish caught in the Eastern Mediterranean Sea. Response is Mercury level (mg/kg). Goal: Treating this as a random sample of all albacore in the area, obtain 95% Confidence Interval for the population mean mercury level.

Fish	1	2	3	4	5	6	7	8	9	10	11	12
Mercury	1.007	1.447	0.763	2.01	1.346	1.243	1.586	0.821	1.735	1.396	1.109	0.993
Fish	13	14	15	16	17	18	19	20	21	22	23	24
Mercury	2.007	1.373	2.242	1.647	1.35	0.948	1.501	1.907	1.952	0.996	1.433	0.866
Fish	25	26	27	28	29	30	31	32	33	34	Mean	StdDev
Mercury	1.049	1.665	2.139	0.534	1.027	1.678	1.214	0.905	1.525	0.763	1.358147	0.440703

$$(1-\alpha) = 0.95 \quad \Rightarrow \quad \alpha = 0.05 \quad n = 34$$

$$\Rightarrow \quad t\left(1-\alpha/2; n-1\right) = t\left(1-0.05/2; 34-1\right) = t\left(0.975; 33\right) = 2.0345$$

$$\overline{Y} = 1.3581 \quad s = 0.4407 \quad \Rightarrow \quad s\left\{\overline{Y}\right\} = \frac{s}{\sqrt{n}} = \frac{0.4407}{\sqrt{34}} = 0.0756$$

$$\overline{Y} \pm t\left(1-\alpha/2; n-1\right) s\left\{\overline{Y}\right\} \quad \equiv \quad 1.3581 \pm 2.0345(0.0756)$$

$$\equiv \quad 1.3581 \pm 0.1538 \quad \equiv \quad \left(1.2043, 1.5119\right)$$

If all possible random samples of size 34 had been obtained, and this calculation had been made for each sample, then 95% of all sample Confidence Intervals would contain the true unknown population mean level μ. Thus we can be 95% confident that μ is between 1.2043 and 1.5119. Note that 90% and 99% Confidence Intervals based on this same sample are as follow (confirm them, and why the lengths differ):

90% Confidence Interval for μ: (1.2302 , 1.4861) 90% Confidence Interval for μ: (1.1516 , 1.5647)

1-Sample *t*-test (2-tailed alternative)

- **2-sided Test: H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$**
- **Decision Rule :**
  - **Conclude $\mu > \mu_0$ if Test Statistic (t^*) > t(1-α/2;n-1)**
  - **Conclude $\mu < \mu_0$ if Test Statistic (t^*) <- t(1-α/2;n-1)**
  - **Do not conclude Conclude $\mu \neq \mu_0$ otherwise**
- ***P*-value: $2P(t(n\text{-}1) \geq |t^*|)$**
- **Test Statistic:**

$$t^* = \frac{\overline{Y} - \mu_0}{s\left\{\overline{Y}\right\}} \qquad s\left\{\overline{Y}\right\} = \frac{s}{\sqrt{n}}$$

1-tailed alternative tests

Upper Tailed $H_0^+ : \mu \leq \mu_0 \quad H_A^+ : \mu > \mu_0$

Decision Rule: Reject H_0^+ if $t^* \geq t\left(1-\alpha; n-1\right)$

P-value: $P\left(t\left(n-1\right) \geq t^*\right)$

Lower Tailed $H_0^- : \mu \geq \mu_0 \quad H_A^- : \mu < \mu_0$

Decision Rule: Reject H_0^- if $t^* \leq -t\left(1-\alpha; n-1\right)$

P-value: $P\left(t\left(n-1\right) \leq t^*\right)$

Note: Tests for μ are generally used when trying to show whether a mean differs from, is above or below some pre-specified value; or when the data are paired differences (such as before/after treatment measures).

Example: The European Union has permissible limit of 1 mg/kg of Mercury in fish. Is μ > 1?

$H_0 : \mu \leq \mu_0 = 1 \quad H_A : \mu > \mu_0 = 1$

$TS : t^* = \frac{\overline{Y} - \mu_0}{s\left\{\overline{Y}\right\}} = \frac{1.3581 - 1}{0.0756} = 4.7836 \geq t\left(0.95; 33\right) = 1.6924 \quad \text{Reject } H_0, \text{ Conclude } \mu > 1$

P-value: $P\left(t(33) \geq 4.7836\right) = .00002$

Comparing 2 Means - Independent Samples

- **Observed individuals/items from the 2 groups are samples from distinct populations (identified by (μ_1,σ_1^2) and (μ_2,σ_2^2))**
- **Measurements across groups are independent**
- **Summary statistics obtained from the 2 groups**

Group 1: Mean: $\overline{Y}$ Std. Dev.: s_1 Sample Size: n_1

Group 2: Mean: $\overline{Z}$ Std. Dev.: s_2 Sample Size: n_2

$$\overline{Y} = \frac{\sum_{i=1}^{n_1} Y_i}{n_1} \qquad s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}\left(Y_i - \overline{Y}\right)^2}{n_1 - 1}}$$ similar calculations for Z

In many settings, we replace $Y_1,...,Y_{n_1}$ with $Y_{11},...,Y_{1n_1}$ and $Z_1,...,Z_{n_2}$ with $Y_{21},...,Y_{2n_2}$

$$\Rightarrow \quad \overline{Y} = \overline{Y}_1 \quad \overline{Z} = \overline{Y}_2$$

Sampling Distribution of $\overline{Y} - \overline{Z}$

- **Underlying distributions normal $\Rightarrow$ sampling distribution is normal, and resulting t-distribution with estimated std. dev.**
- **Mean, variance, standard error (Std. Dev. of estimator)**

$$E\left\{\overline{Y} - \overline{Z}\right\} = \mu_{\overline{Y}-\overline{Z}} = \mu_1 - \mu_2$$

$$\sigma^2\left\{\overline{Y} - \overline{Z}\right\} = \sigma^2_{\overline{Y}-\overline{Z}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \qquad \sigma_{\overline{Y}-\overline{Z}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\sigma_1^2 = \sigma_2^2 \quad \Rightarrow \quad \frac{\left(\overline{Y} - \overline{Z}\right) - \left(\mu_1 - \mu_2\right)}{s\left\{\overline{Y} - \overline{Z}\right\}} \sim t \text{ with df} = n_1 + n_2 - 2$$

$$\text{where: } s\left\{\overline{Y} - \overline{Z}\right\} = s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad s^2 = \frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{n_1 + n_2 - 2}$$

Inference for $\mu_1 - \mu_2$ - Normal Populations – Equal variances

$$(1-\alpha)100\% \text{ Confidence Interval:}$$
$$\left(\overline{Y}-\overline{Z}\right) \pm t\left(1-\alpha/2; n_1+n_2-2\right) s\left\{\overline{Y}-\overline{Z}\right\}$$

- **Interpretation (at the α significance level):**
  - **If interval contains 0, do not reject H_0: $\mu_1 = \mu_2$**
  - **If interval is strictly positive, conclude that $\mu_1 > \mu_2$**
  - **If interval is strictly negative, conclude that $\mu_1 < \mu_2$**

$$H_0: \mu_1 - \mu_2 = 0 \qquad H_A: \mu_1 - \mu_2 \neq 0$$
$$\text{Test Stat}: t^* = \frac{\overline{Y}-\overline{Z}}{s\left\{\overline{Y}-\overline{Z}\right\}}$$
$$\text{Reject Reg}: \left|t^*\right| \geq t\left(1-\alpha/2; n_1+n_2-2\right)$$

Example – Children's Participation in Meal Preparation and Caloric Intake

Experiment had 2 conditions: Child participated in Cooking Meal, and Parent only cooking meal. Response measured: Total Energy Intake (kcals). Total of 47 participants: 25 in Child cooks (Y), 22 in Parent cooks (Z).

Child Cooks: $\overline{Y} = 431.4 \quad s_1 = 105.7 \quad n_1 = 25$ Parent Cooks: $\overline{Z} = 346.8 \quad s_2 = 99.5 \quad n_2 = 22$

$$\overline{Y}-\overline{Z} = 431.4 - 346.8 = 84.6 \qquad t\left(1-.05/2; 25+22-2\right) = t\left(.975; 45\right) = 2.0141$$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(25-1)105.7^2 + (22-1)99.5^2}{25+22-2} = \frac{476045}{45} = 10578.78 \quad \Rightarrow \quad s = 102.8532$$

$$s\left\{\overline{Y}-\overline{Z}\right\} = s\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 102.8532\sqrt{\left(\frac{1}{25}+\frac{1}{22}\right)} = 102.8532\left(0.2923\right) = 30.0667$$

95% CI for $\mu_1 - \mu_2$: $\left(\overline{Y}-\overline{Z}\right) \pm t\left(.975;45\right) s\left\{\overline{Y}-\overline{Z}\right\} \equiv 84.6 \pm 2.0141(30.0667) \equiv 84.6 \pm 60.6 \equiv \left(24.0, 145.2\right)$

Testing: $H_0: \mu_1 - \mu_2 = 0$ vs $H_A: \mu_1 - \mu_2 \neq 0$

$$TS: t^* = \frac{\overline{Y}-\overline{Z}}{s\left\{\overline{Y}-\overline{Z}\right\}} = \frac{84.6}{30.0667} = 2.8137 > t\left(.975;45\right) = 2.0141 \qquad P = 2P\left(t(45) \geq 2.8137\right) = 2(.0036) = .0072$$

Sampling Distribution of s^2 (Normal Data)

- Population variance (σ^2) is a fixed (unknown) parameter based on the population of measurements
- Sample variance (s^2) varies from sample to sample (just as sample mean does)
- When $Y \sim N(\mu,\sigma^2)$, the distribution of (a multiple of) s^2 is Chi-Square with n-1 degrees of freedom. Unlike inference on means, the normality assumption is very important.
- $(n-1)s^2/\sigma^2 \sim \chi^2$ with df=n-1

(1-α)100% Confidence Interval for σ^2 (or σ)

- Step 1: Obtain a random sample of n items from the population, compute s^2
- Step 2: Obtain χ^2_L and χ^2_U from table of critical values for chi-square distribution with n-1 df
- Step 3: Compute the confidence interval for σ^2 based on the formula below and take square roots of bounds for σ^2 to obtain confidence interval for σ

$$(1-\alpha)100\% \text{ CI for } \sigma^2: \quad \left(\frac{(n-1)s^2}{\chi^2_U}, \frac{(n-1)s^2}{\chi^2_L}\right)$$

$$\text{where: } \chi^2_U = \chi^2\left(1-\alpha/2; n-1\right) \qquad \chi^2_L = \chi^2\left(\alpha/2; n-1\right)$$

Example: Mercury Levels in Albacore Fish from Eastern Mediterranean (Continued)

$$(1-\alpha) = 0.95 \quad \Rightarrow \quad \alpha = 0.05 \quad \Rightarrow \quad \alpha/2 = 0.025 \quad \Rightarrow \quad 1-\alpha/2 = 0.975$$

$$n = 34 \quad \Rightarrow \quad \chi^2_U = \chi^2\left(1-\alpha/2; n-1\right) = \chi^2\left(.975; 33\right) = 50.73 \qquad \chi^2_L = \chi^2\left(.025; 33\right) = 19.05$$

$$s = 0.4407 \quad \Rightarrow \quad s^2 = 0.4407^2 = 0.1942 \quad \Rightarrow \quad (n-1)s^2 = 33(0.1942) = 6.4092$$

$$(1-\alpha)100\% \text{ CI for } \sigma^2: \quad \left(\frac{(n-1)s^2}{\chi^2_U}, \frac{(n-1)s^2}{\chi^2_L}\right) \quad \equiv \quad \left(\frac{6.4092}{50.73}, \frac{6.4092}{19.05}\right) \quad \equiv \quad (0.1263, 0.3364)$$

$$(1-\alpha)100\% \text{ CI for } \sigma: \quad \left(\sqrt{0.1263}, \sqrt{0.3364}\right) \quad \equiv \quad (0.3364, 0.5800)$$

Statistical Test for σ^2

- **Null and alternative hypotheses**
  - 1-sided (upper tail): H_0: $\sigma^2 \leq \sigma_0^2$ H_a: $\sigma^2 > \sigma_0^2$
  - 1-sided (lower tail): H_0: $\sigma^2 \geq \sigma_0^2$ H_a: $\sigma^2 < \sigma_0^2$
  - 2-sided: H_0: $\sigma^2 = \sigma_0^2$ H_a: $\sigma^2 \neq \sigma_0^2$
- **Test Statistic**

$$\boxed{\chi^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}}$$

- **Decision Rule based on chi-square distribution w/ df=*n*-1:**
  - 1-sided (upper tail): Reject H_0 if $\chi_{obs}^2 > \chi_U^2 = \chi^2(1-\alpha;n-1)$
  - 1-sided (lower tail): Reject H_0 if $\chi_{obs}^2 < \chi_L^2 = \chi^2(\alpha;n-1)$
  - 2-sided: Reject H_0 if $\chi_{obs}^2 < \chi_L^2 = \chi^2(\alpha/2;n-1)$(Conclude $\sigma^2 < \sigma_0^2$) or if $\chi_{obs}^2 > \chi_U^2 = \chi^2(1-\alpha/2;n-1)$ (Conclude $\sigma^2 > \sigma_0^2$)

There are not too many practical cases where there is a null value to test, except in cases where firms may need to demonstrate that variation in purity of a chemical or compound is below some nominal level, or that variation in measurements of manufactured parts is below some nominal level.

Note that most decisions can be obtained based on the confidence interval for the population variance (or standard deviation).

Inferences Regarding 2 Population Variances

- **Goal: Compare variances between 2 populations**
- **Parameter:** $\frac{\sigma_1^2}{\sigma_2^2}$ **(Ratio is 1 when variances are equal)**
- **Estimator:** $\frac{s_1^2}{s_2^2}$ **(Ratio of sample variances)**
- **Distribution of (multiple) of estimator (Normal Data):**

$$\boxed{\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{s_1^2 / s_2^2}{\sigma_1^2 / \sigma_2^2} \sim F \quad \text{with df}_1 = n_1 - 1 \quad \text{and df}_2 = n_2 - 1}$$

Test Comparing Two Population Variances

1-Sided Test: $H_0 : \sigma_1^2 \leq \sigma_2^2 \quad H_a : \sigma_1^2 > \sigma_2^2$

Test Statistic: $F_{obs} = \frac{s_1^2}{s_2^2}$ Rejection Region: $F_{obs} \geq F\left(1-\alpha; n_1 - 1, n_2 - 1\right)$ $P-$value: $P(F \geq F_{obs})$

2-Sided Test: $H_0 : \sigma_1^2 = \sigma_2^2 \quad H_a : \sigma_1^2 \neq \sigma_2^2$

Test Statistic: $F_{obs} = \frac{s_1^2}{s_2^2}$

Rejection Region: $F_{obs} \geq F\left(1-\alpha/2; n_1 - 1, n_2 - 1\right) \quad (\sigma_1^2 > \sigma_2^2)$

or $F_{obs} \leq F\left(\alpha/2; n_1 - 1, n_2 - 1\right) = 1/F\left(1-\alpha/2; n_2 - 1, n_1 - 1\right) \quad (\sigma_1^2 < \sigma_2^2)$

$P-$value: $2\min(P(F \geq F_{obs}), P(F \leq F_{obs}))$

(1-α)100% Confidence Interval for σ_1^2/σ_2^2

- Obtain ratio of sample variances $s_1^2/s_2^2 = (s_1/s_2)^2$
- Choose α, and obtain:
  - $F_L = F(\alpha/2, \text{n1-1, n2-1}) = 1/ F(1-\alpha/2, \text{n2-1, n1-1})$
  - $F_U = F(1-\alpha/2, \text{n1-1, n2-1})$
- Compute Confidence Interval:

$$\left[\frac{s_1^2}{s_2^2} F_L \ , \ \frac{s_1^2}{s_2^2} F_U \right]$$

Conclude population variances unequal if interval does not contain 1

Example – Children's Participation in Meal Preparation and Caloric Intake (Continued)

2-Sided Test: $H_0 : \sigma_1^2 = \sigma_2^2 \quad H_a : \sigma_1^2 \neq \sigma_2^2$

Test Statistic: $F_{obs} = \frac{s_1^2}{s_2^2} = \frac{105.7^2}{99.5^2} = 1.13$

Rejection Region: $F_{obs} \geq F\left(1-\alpha/2; n_1-1, n_2-1\right) = F\left(1-.025; 25-1, 22-1\right) = F\left(.975; 24, 21\right) = 2.3675 \quad (\sigma_1^2 > \sigma_2^2)$

or $F_{obs} \leq F\left(.025; 24, 21\right) = 1/F\left(.975; 21, 24\right) = 0.4327 \quad (\sigma_1^2 < \sigma_2^2)$

$P-$value: $2\min(P(F \geq F_{obs}), P(F \leq F_{obs})) = 2\min\left(.3912, .6088\right) = 0.7824$

95% Confidence Interval for $\frac{\sigma_1^2}{\sigma_2^2}$

$F_L = F\left(.025; 24, 21\right) = 1/F\left(.975; 21, 24\right) = 0.4327$

$F_U = F\left(.975; 24, 21\right) = 2.3675$

$$\left[\frac{s_1^2}{s_2^2} F_L , \frac{s_1^2}{s_2^2} F_U\right] \equiv \left[1.13(0.4327) , 1.13(2.3675)\right] \equiv \left[0.49 , 2.68\right]$$

What do you conclude?

Data Sources:

New York City Street Café's:

https://nycopendata.socrata.com/Business/Sidewalk-Cafes/6k68-kc8u

Women's Professional Soccer:

http://www.nwslsoccer.com/

Irish Premier League Soccer:

www.**soccerpunter**.com/

Mercury Levels in Albacore:

S. Mol, O. Ozden, S. Karakulak (2012). "Levels of Selected Metals in Albacore (Thunnus alalunga, Bonaterre, 1788) from the Eastern Mediterranean, *Journal of Aquatic Food Product Technology*, Vol. 21, #2, pp. 111-117.

Children/Parent Cooking Effects on Food Intake:

K. van der Horst, A. Ferrage, A. Rytz (2014). "Involving Children in Meal Preparation: Effects on Food Intake," *Appetite*, Vol. 79, pp. 18-24.

Chapter 1 – Linear Regression with 1 Predictor

Statistical Model

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,\ldots,n$

where:

- Y_i is the (random) response for the i^{th} case
- β_0, β_1 are parameters
- X_i is a known constant, the value of the predictor variable for the i^{th} case
- ε_i is a random error term, such that: $E\{\varepsilon_i\} = 0 \quad \sigma^2\{\varepsilon_i\} = \sigma^2 \quad \sigma\{\varepsilon_i, \varepsilon_j\} = 0 \quad \forall\, i, j \ni i \neq j$

The last point states that the random errors are independent (uncorrelated), with mean 0, and variance σ^2. This also implies that:

$E\{Y_i\} = \beta_0 + \beta_1 X_i \qquad \sigma^2\{Y_i\} = \sigma^2 \quad \sigma\{Y_i, Y_j\} = 0$

Thus, β_0 represents the mean response when $X = 0$ (assuming that is reasonable level of X), and is referred to as the ***Y*-intercept**. Also, β_1 represent the change in the mean response as X increases by 1 unit, and is called the **slope**.

Least Squares Estimation of Model Parameters

In practice, the parameters β_0 and β_1 are unknown and must be estimated. One widely used criterion is to minimize the error sum of squares:

$$\boxed{\begin{aligned} &Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Rightarrow \quad \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) \\ &Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \end{aligned}}$$

This is done by calculus, by taking the partial derivatives of Q with respect to β_0 and β_1 and setting each equation to 0. The values of β_0 and β_1 that set these equations to 0 are the **least squares estimates** and are labelled b_0 and b_1.

First, take the partial derivatives of Q with respect to β_0 and β_1:

$$\boxed{\begin{aligned} &\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))(-1) \qquad (1) \\ &\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))(-X_i) \qquad (2) \end{aligned}}$$

Next, set these 2 equations to 0, replacing β_0 and β_1 with b_0 and b_1 since these are the values that minimize the error sum of squares:

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i \qquad (1a)$$

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)X_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (2a)$$

These two equations are referred to as the **normal equations** (although, note that we have said nothing YET, about normally distributed data).

Solving these two equations yields:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} Y_i = \sum_{i=1}^{n} k_i Y_i$$

$$b_0 = \overline{Y} - b_1 \overline{X} = \sum_{i=1}^{n} \left[\frac{1}{n} - \overline{X} k_i \right] Y_i = \sum_{i=1}^{n} l_i Y_i$$

where k_i and l_i are constants, and Y_i is a random variable with mean and variance given above:

$$k_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad l_i = \frac{1}{n} - \overline{X} k_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

The **fitted regression line**, also known as the **prediction equation** is:

$$\hat{Y} = b_0 + b_1 X$$

The **fitted values** for the individual observations are obtained by plugging in the corresponding level of the predictor variable (X_i) into the fitted equation. The **residuals** are the vertical distances between the **observed values** (Y_i) and their **fitted values** ($\hat{Y}_i$), and are denoted as e_i.

$$\hat{Y}_i = b_0 + b_1 X_i \qquad\qquad e_i = Y_i - \hat{Y}_i$$

Properties of the fitted regression line

- $\sum_{i=1}^{n} e_i = 0$ The residuals sum to 0
- $\sum_{i=1}^{n} X_i e_i = 0$ The sum of the weighted (by X) residuals is 0
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ The sum of the weighted (by $\hat{Y}$) residuals is 0
- The regression line goes through the point $(\overline{X}, \overline{Y})$

These can be derived via their definitions and the normal equations:

$$\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} \left(b_0 + b_1 X_i\right) = nb_0 + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \quad \Rightarrow \quad \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n} e_i = 0 \qquad (1a)$$

$$\sum_{i=1}^{n} X_i \hat{Y}_i = \sum_{i=1}^{n} X_i \left(b_0 + b_1 X_i\right) = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \quad \Rightarrow \quad \sum_{i=1}^{n} X_i \left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n} X_i e_i = 0 \qquad (2a)$$

Estimation of the Error Variance

Note that for a random variable, its variance is the expected value of the squared deviation from the mean. That is, for a random variable W, with mean μ_W its variance is:

$$\sigma^2\{W\} = E\{(W - \mu_W)^2\}$$

For the simple linear regression model, the errors have mean 0, and variance σ^2. This means that for the actual observed values Y_i, their mean and variance are as follows:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \qquad \sigma^2\{Y_i\} = E\left\{\left(Y_i - \left(\beta_0 + \beta_1 X_i\right)\right)^2\right\} = \sigma^2$$

First, we replace the unknown mean $\beta_0 + \beta_1 X_i$ with its fitted value $\hat{Y}_i = b_0 + b_1 X_i$, then we take the "average" squared distance from the observed values to their fitted values. We divide the sum of squared errors by *n*-2 to obtain an unbiased estimate of σ^2 (recall how you computed a sample variance when sampling from a single population).

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Common notation is to label the numerator as the **error sum of squares (SSE)**.

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Also, the estimated variance is referred to as the **error (or residual) mean square (MSE)**.

$$\boxed{MSE \quad = \quad s^2 \quad = \quad \frac{SSE}{n-2}}$$

To obtain an estimate of the standard deviation (which is in the units of the data), we take the square root of the error mean square. $s = \sqrt{MSE}$.
A shortcut formula for the error sum of squares, which can cause problems due to round-off errors is:

$$\boxed{SSE \quad = \quad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - b_1\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}$$

Some notation makes life easier when writing out elements of the regression model:

$$SS_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n} = \sum_{i=1}^{n}X_i^2 - n\left(\overline{X}\right)^2$$

$$SS_{XY} = \sum_{i=1}^{n}\left[(X_i - \overline{X})(Y_i - \overline{Y})\right] = \sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n} = \sum_{i=1}^{n}X_iY_i - n\overline{X}\,\overline{Y}$$

$$SS_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}Y_i^2 - \frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n} = \sum_{i=1}^{n}Y_i^2 - n\left(\overline{Y}\right)^2$$

Note that we will be able to obtain most all of the simple linear regression analysis from these quantities, the sample means, and the sample size.

$$b_1 = \frac{SS_{XY}}{SS_{XX}} \qquad b_0 = \overline{Y} - b_1\overline{X} \qquad SSE = SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}} = SS_{YY} - b_1 SS_{XY} \qquad s^2 = MSE = \frac{SSE}{n-2}$$

<u>Normal Error Regression Model (Assumes STA 4322)</u>

If we add further that the random errors follow a normal distribution, then the response variable also has a normal distribution, with mean and variance given above. The notation, we will use for the errors, and the data is:

$$\varepsilon_i \sim N(0, \sigma^2) \qquad Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

The density function for the i^{th} observation is:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

The likelihood function, is the product of the individual density functions (due to the independence assumption on the random errors).

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 \right]$$

The values of $\beta_0, \beta_1, \sigma^2$ that maximize the likelihood function are referred to as **maximum likelihood estimators**. The MLE's are denoted as: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_2$. Note that the natural logarithm of the likelihood is maximized by the same values of $\beta_0, \beta_1, \sigma^2$ that maximize the likelihood function, and it's easier to work with the log likelihood function.

$$\log_e L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Taking partial derivatives with respect to $\beta_0, \beta_1, \sigma^2$ yields:

$$\frac{\partial \log L}{\partial \beta_0} = -2\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-1) \qquad (4)$$

$$\frac{\partial \log L}{\partial \beta_1} = -2\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-X_i) \qquad (5)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 \qquad (6)$$

Setting these three equations to 0, and placing "hats" on parameters denoting the maximum likelihood estimators, we get the following three equations:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \qquad (4a)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad (5a)$$

$$\frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{n}{\hat{\sigma}^2} \qquad (6a)$$

From equations 4*a* and 5*a*, we see that the maximum likelihood estimators are the same as the least squares estimators (these are the normal equations). However, from equation 6*a*, we obtain the maximum likelihood estimator for the error variance as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$$

This estimator is biased downward. We will use the unbiased estimator $s^2 = MSE$ throughout this course to estimate the error variance.

<u>Example – U.S. State Non-Fuel Mineral Production vs Land Area (2011).</u>

Non-Fuel mineral production ($10M) and land area (1000m^2) for the 50 United States in 2011.

Source: http://minerals.er.usgs.gov/minerals/pubs/commodity/statistical_summary/index.html#myb (retrieved 6/23/2014).

The following EXCEL spreadsheet gives the data in a form that is easier to read. The original data are in an EXCEL file in Columns A-C and Rows 1-51 (variable names in row 1, numeric data in rows 2-51). Note that Column A contains the state postal abbreviation, B contains Area, and C contains mineral production.

state	Area	Mineral	state	Area	Mineral	state	Area	Mineral	state	Area	Mineral	state	Area	Mineral
AL	50.74	96.0	HI	6.42	10.1	MA	7.84	22.5	NM	121.36	125.0	SD	75.89	31.2
AK	567.40	381.0	ID	82.75	132.0	MI	58.11	241.0	NY	47.21	134.0	TN	41.22	87.8
AZ	113.64	839.0	IL	55.58	107.0	MN	79.61	449.0	NC	48.71	84.3	TX	261.80	303.0
AR	52.07	78.9	IN	35.87	76.2	MS	46.91	19.5	ND	68.98	12.5	UT	82.14	430.0
CA	155.96	321.0	IA	55.87	65.3	MO	68.89	220.0	OH	40.95	96.2	VT	9.25	11.8
CO	103.72	193.0	KS	81.82	112.0	MT	145.55	144.0	OK	68.67	60.8	VA	39.59	119.0
CT	4.85	15.6	KY	39.73	79.1	NE	76.87	23.8	OR	96.00	30.5	WA	66.54	74.2
DE	1.95	1.1	LA	43.56	46.5	NV	109.83	1000.0	PA	44.82	160.0	WV	24.23	32.4
FL	53.93	343.0	ME	30.86	11.8	NH	8.97	10.0	RI	1.05	4.2	WI	54.31	68.3
GA	57.91	145.0	MD	9.77	29.3	NJ	7.42	27.5	SC	30.11	48.3	WY	97.11	214.0

Which variable is more likely to "cause" the other variable?

AREA → MINERAL or **MINERAL → AREA**

While we will use R for statistical analyses this semester that would be way too time consuming (if even possible) in EXCEL, EXCEL does have some nice built-in functions to make calculations on ranges of cells.

- =COUNT(*range*) - Computes the number of values in the range
- =SUM(*range*) - Computes the sum for the values in the range
- =AVERAGE(*range*) - Computes the sample mean for the values in the range
- =VAR(*range*) - Computes the sample mean for the values in the range
- =STDEV(*range*) - Computes the sample mean for the values in the range
- =SUMSQ(*range*) - Computes the sum of squares for the values in the range
- =DEVSQ(*range*) - Computes the sum of squared deviations from the mean
- =SUMPRODUCT(*range1*,*range2*) - Computes the sum of products of each pair of elements of 2 ranges of equal length
- =COVAR(*range1*,*range2*) - Computes the covariance of two ranges of equal length, using n as the denominator, not n-1. In later versions, =COVARIANCE.S(*range1*,*range2*) is available, using n-1.

Making use of these, we can "brute-force" obtain the estimated regression equation and estimated error variance. First, obtain the means and sums of squares and cross-products needed to obtain the regression equation.

$n:\quad =\text{COUNT}(\text{B2:B51})\qquad \sum_{i=1}^{n} X_i:\quad =\text{SUM}(\text{B2:B51})\qquad \sum_{i=1}^{n} Y_i:\quad =\text{SUM}(\text{C2:C51})$

$\overline{X}:\quad =\text{AVERAGE}(\text{B2:B51})\qquad \overline{Y}:\quad =\text{AVERAGE}(\text{C2:C51})\qquad \sum_{i=1}^{n} X_i^2:\quad =\text{SUMSQ}(\text{B2:B51})$

$\sum_{i=1}^{n} Y_i^2:\quad =\text{SUMSQ}(\text{C2:C51})\qquad \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2:\quad =\text{DEVSQ}(\text{B2:B51})\qquad \sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2:\quad =\text{DEVSQ}(\text{C2:C51})$

$\sum_{i=1}^{n} X_iY_i:\quad =\text{SUMPRODUCT}(\text{B2:B51,C2:C51})\qquad \frac{1}{n}\sum_{i=1}^{n}\left[\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)\right]:\quad =\text{COVAR}(\text{B2:B51,C2:C51})$

n	**50.00**	**sum(Y^2)**	**2975248.32**
X-bar	**70.69**	**sum(XY)**	**856554.66**
Y-bar	**147.35**	**SS_XX**	**357703.85**
sum(X)	**3534.29**	**SS_YY**	**1889585.31**
sum(Y)	**7367.71**	**COV(X,Y)**	**6715.24**
sum(X^2)	**607528.11**	**SS_XY**	**335762.03**

Note that when using formulas with "multiple steps" you will find there are "small" rounding errors.

$$SS_{XX}=\sum_{i=1}^{n}(X_i-\overline{X})^2=357703.85$$

$$=\sum_{i=1}^{n}X_i^2-\frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}=607528.1-\frac{(3534.29)^2}{50}=\underline{\qquad\qquad\qquad\qquad}$$

$$=\sum_{i=1}^{n}X_i^2-n\left(\overline{X}\right)^2=607528.1-50(70.69)^2=\underline{\qquad\qquad\qquad\qquad}$$

$$SS_{XY}=\sum_{i=1}^{n}\left[(X_i-\overline{X})(Y_i-\overline{Y})\right]=n\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left[(X_i-\overline{X})(Y_i-\overline{Y})\right]=50\left(6715.24\right)=335762$$

$$=\sum_{i=1}^{n}X_iY_i-\frac{\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n}=856554.66-\frac{(3534.29)(7367.71)}{50}=\underline{\qquad\qquad\qquad\qquad}$$

$$=\sum_{i=1}^{n}X_iY_i-n\overline{X}\,\overline{Y}=856554.66-50(70.69)(147.35)=\underline{\qquad\qquad\qquad\qquad}$$

$$SS_{YY}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2=1889585.31$$

$$=\sum_{i=1}^{n}Y_i^2-\frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}=2975248.32-\frac{(7367.71)^2}{50}=\underline{\qquad\qquad\qquad\qquad}$$

$$\sum_{i=1}^{n}Y_i^2-n\left(\overline{Y}\right)^2=2975248.32-50(147.35)^2=\underline{\qquad\qquad\qquad\qquad}$$

Next compute the estimated regression coefficients, fitted equation, and estimated error variance and standard deviation.

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{335762}{357703.85} = 0.9387 \qquad b_0 = \overline{Y} - b_1\overline{X} = 147.35 - 0.9387(70.69) = 80.9933$$

$\hat{Y} = 80.99 + 0.94X$ or using symbols better related to data: $\hat{M} = 80.99 + 0.94A$

For State 1 (Alabama): Area = $X_1 =$ 50.74 and Mineral = $Y_1 = 96.0$

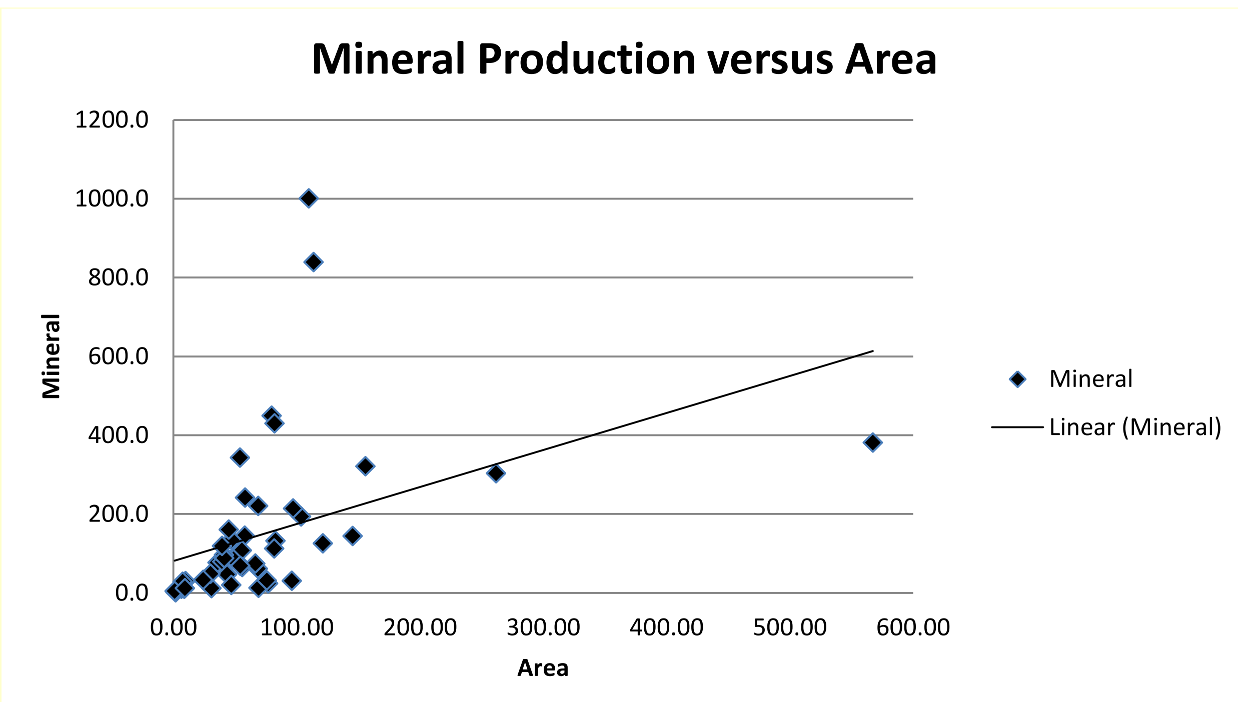
$$\Rightarrow \quad \hat{Y}_1 = b_0 + b_1 X_1 = 80.99 + 0.94(50.74) = 80.99 + 47.70 = 128.69 \quad \Rightarrow \quad e_1 = Y_1 - \hat{Y}_1 = 96.0 - 128.69 = -32.69$$

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$$

$$= SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}} = 1889585.31 - \frac{(335762)^2}{357703.85} = 1889585.31 - 315166.08 = 1574419.23$$

$$\Rightarrow \quad s^2 = MSE = \frac{SSE}{n-2} = 32800.40 \quad \Rightarrow \quad s = \sqrt{32800.40} = 181.11$$

A plot of the data and the fitted equation are given below, obtained from EXCEL.



As land area increases by 1 unit (1000 mile^2), mineral value increases on average by 0.94 units (\$10M). The intercept has no physical meaning, as no states have an area of 0.

Note that while there is a tendency for larger states to have higher mineral production, there are many states that the line does not fit well for. This issue among others will be considered in later chapters, and a model with both variables log transformed is fit below.

EXCEL (Using Built-in Data Analysis Package)

Regression Coefficients (and standard errors/t-tests/CI's, which will be covered in Chapter 2)

	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	81.004	33.379	2.427	0.0190	13.891	148.118
Area	0.939	0.303	3.100	0.0032	0.330	1.548

Data Cells, Fitted Values and Residuals (Copied and Pasted to fit better on page)

state	Area	Mineral	Fitted	Residual		state	Area	Mineral	Fitted	Residual
AL	50.74	96.0	128.6356	-32.6356		MT	145.55	144.0	217.628	-73.628
AK	567.40	381.0	613.5996	-232.6		NE	76.87	23.8	153.1609	-129.361
AZ	113.64	839.0	187.6688	651.3312		NV	109.83	1000.0	184.0935	815.9065
AR	52.07	78.9	129.8784	-50.9784		NH	8.97	10.0	89.4222	-79.4522
CA	155.96	321.0	227.3967	93.60334		NJ	7.42	27.5	87.96634	-60.4663
CO	103.72	193.0	178.3602	14.63984		NM	121.36	125.0	194.9162	-69.9162
CT	4.85	15.6	85.5521	-69.9521		NY	47.21	134.0	125.3222	8.677841
DE	1.95	1.1	82.83844	-81.7184		NC	48.71	84.3	126.7273	-42.4273
FL	53.93	343.0	131.6234	211.3766		ND	68.98	12.5	145.7493	-133.249
GA	57.91	145.0	135.3583	9.641697		OH	40.95	96.2	119.4405	-23.2405
HI	6.42	10.1	87.03331	-76.9333		OK	68.67	60.8	145.4592	-84.6592
ID	82.75	132.0	158.6755	-26.6755		OR	96.00	30.5	171.1128	-140.613
IL	55.58	107.0	133.1787	-26.1787		PA	44.82	160.0	123.0722	36.92781
IN	35.87	76.2	114.6712	-38.4712		RI	1.05	4.2	81.9852	-77.7652
IA	55.87	65.3	133.4463	-68.1463		SC	30.11	48.3	109.2664	-60.9664
KS	81.82	112.0	157.8007	-45.8007		SD	75.89	31.2	152.2345	-121.034
KY	39.73	79.1	118.2954	-39.1954		TN	41.22	87.8	119.693	-31.893
LA	43.56	46.5	121.8942	-75.3942		TX	261.80	303.0	326.7425	-23.7425
ME	30.86	11.8	109.9732	-98.1732		UT	82.14	430.0	158.1095	271.8905
MD	9.77	29.3	90.17876	-60.8788		VT	9.25	11.8	89.6869	-77.8869
MA	7.84	22.5	88.36339	-65.8634		VA	39.59	119.0	118.1696	0.830425
MI	58.11	241.0	135.5498	105.4502		WA	66.54	74.2	143.4664	-69.2664
MN	79.61	449.0	155.731	293.269		WV	24.23	32.4	103.748	-71.348
MS	46.91	19.5	125.034	-105.534		WI	54.31	68.3	131.9829	-63.6829
MO	68.89	220.0	145.6648	74.33522		WY	97.11	214.0	172.1528	41.84719

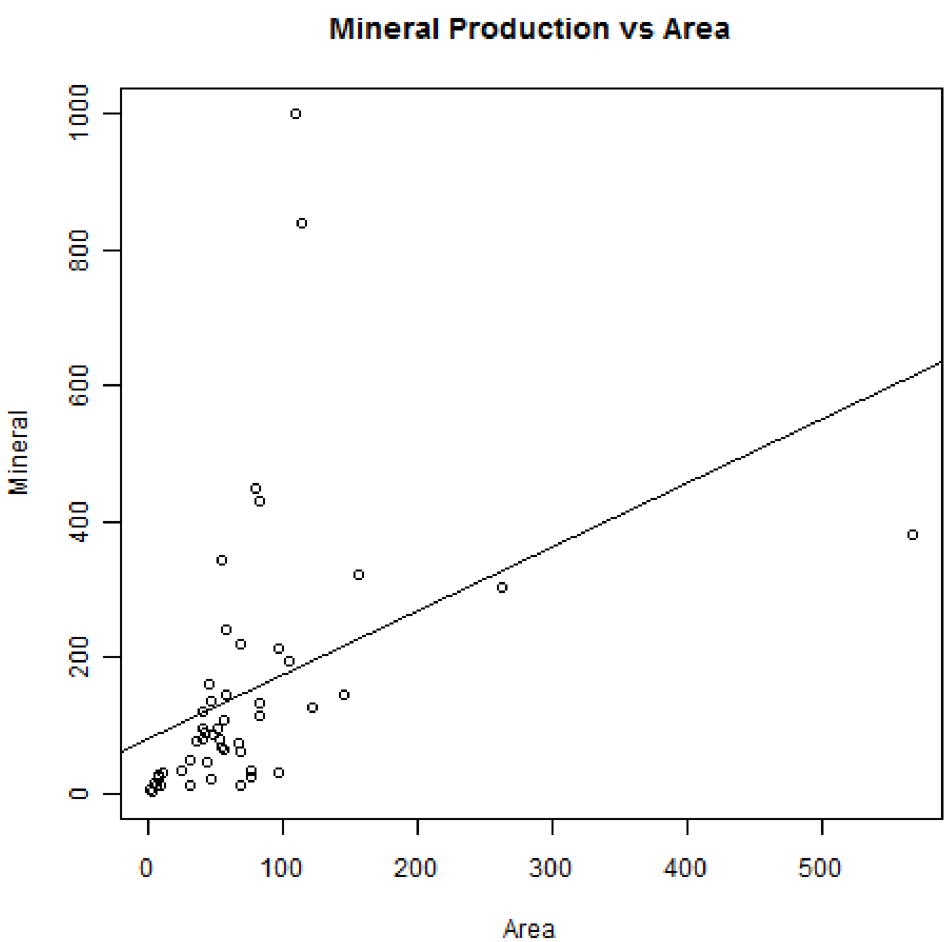
R Program for Regression Analysis and Plot

```
png("F:\\blue_drive\\Rmisc\\graphs\\mineral1.png")

mineral1 <- read.table("http://www.stat.ufl.edu/~winner/sta4210/mydata/mineral1.txt",
        header=T)

attach(mineral1)

min.reg1 <- lm(Mineral ~ Area)
summary(min.reg1)

plot(Area,Mineral,xlab="Area",ylab="Mineral",main="Mineral Production vs Area")
abline(min.reg1)

dev.off()
```

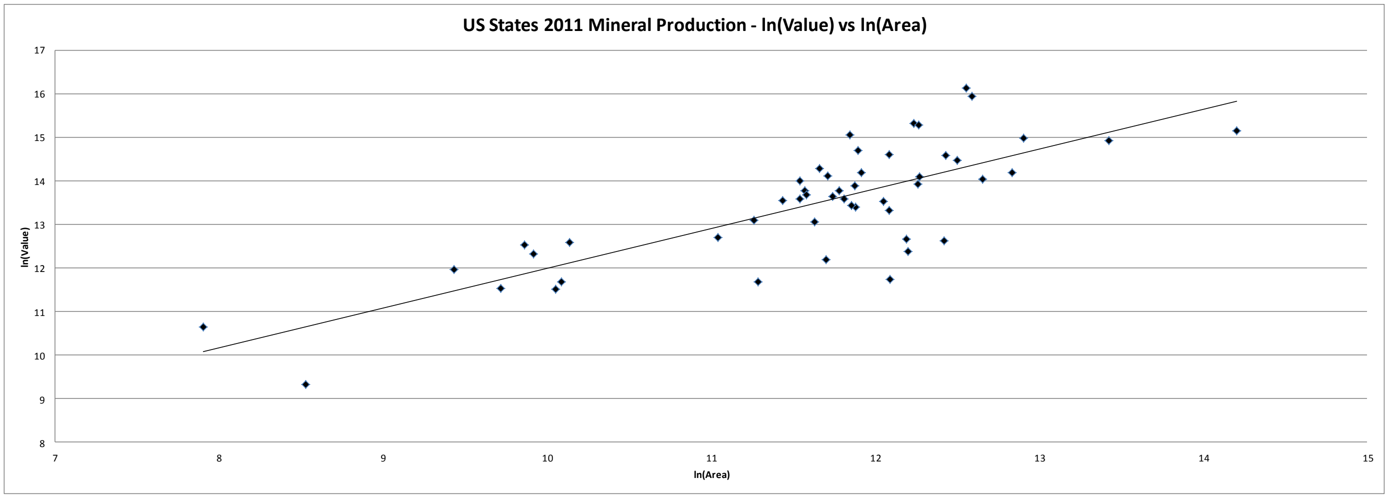
R Regression Output:

```
Call:
lm(formula = Mineral ~ Area)

Residuals:
    Min      1Q  Median      3Q     Max
-232.60  -76.54  -55.72    6.72  815.90

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.0023    33.3793   2.427  0.01904 *
Area          0.9387     0.3028   3.100  0.00324 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 181.1 on 48 degrees of freedom
Multiple R-squared: 0.1668,     Adjusted R-squared: 0.1494
F-statistic: 9.609 on 1 and 48 DF,  p-value: 0.003236
```

R Graphics Output:



Analysis when each variable has been transformed by taking natural logarithms:



Note that the linear relation appears to fit much better when both of these highly skewed variables are log transformed.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.888	1.221	2.366	0.0221	0.434	5.342
lnAREA	0.911	0.105	8.708	0.0000	0.701	1.121

As ln(AREA) increases 1 unit, ln(VALUE) increases by 0.911 units.

Note: When both variables are log transformed the physical meaning of the slope represents percent changes in variables in their original units. In this case, we would say that a **1 percent increase in area is associated with a 0.911 percent change in mineral production value**.

Example – LSD Concentration and Math Scores

A pharmacodynamic study was conducted at Yale in the 1960's to determine the relationship between LSD concentration and math scores in a group of volunteers. The independent (predictor) variable was the mean tissue concentration of LSD in a group of 5 volunteers, and the dependent (response) variable was the mean math score among the volunteers. There were *n*=7 observations, collected at different time points throughout the experiment.

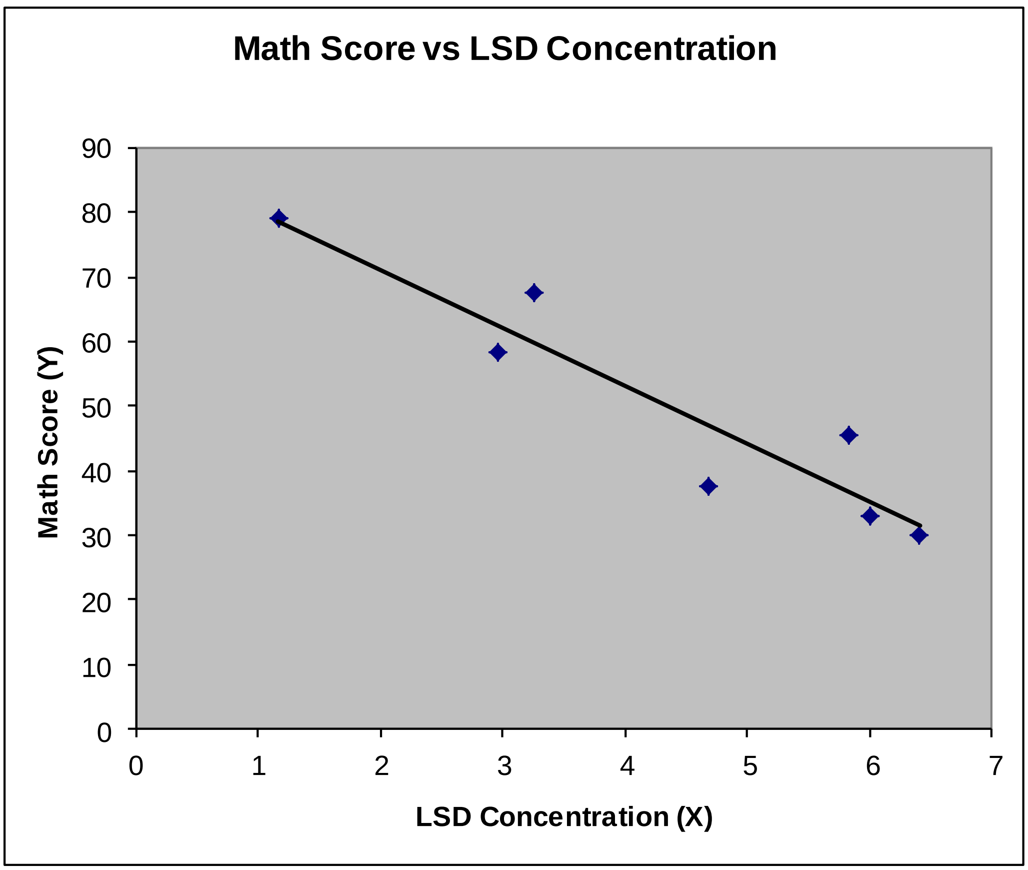
Source: Wagner, J.G., Agahajanian, G.K., and Bing, O.H. (1968), "Correlation of Performance Test Scores with Tissue Concentration of Lysergic Acid Diethylamide in Human Subjects," *Clinical Pharmacology and Therapeutics*, 9:635-638.

The following EXCEL spreadsheet gives the data and all pertinent calculations in spreadsheet form.

Time (*i*)	Score (*Y*)	Conc (*X*)	*Y-Ybar*	*X-Xbar*	*(Y-Ybar)**2*	*(X-Xbar)**2*	*(X-Xbar)(Y-Ybar)*	*Yhat*	*e*	e**2
1	78.93	1.17	28.84286	-3.162857	831.910408	10.0036653	-91.2258367	78.5828	0.3472	0.1205
2	58.2	2.97	8.112857	-1.362857	65.818451	1.85737959	-11.0566653	62.36576	-4.1658	17.354
3	67.47	3.26	17.38286	-1.072857	302.163722	1.15102245	-18.6493225	59.75301	7.717	59.552
4	37.47	4.69	-12.61714	0.357143	159.192294	0.12755102	-4.50612245	46.86948	-9.3995	88.35
5	45.65	5.83	-4.437143	1.497143	19.6882367	2.24143673	-6.64303674	36.59868	9.0513	81.926
6	32.92	6	-17.16714	1.667143	294.710794	2.77936531	-28.6200796	35.06708	-2.1471	4.6099
7	29.97	6.41	-20.11714	2.077143	404.699437	4.31452245	-41.7861796	31.37319	-1.4032	1.969
Sum	350.61	30.33	0	0	2078.18334	22.4749429	-202.487243	350.61	1.00E-14	253.88
Mean	50.0871429	4.3328571								
b1	-9.009466									
b0	89.123874									
MSE	50.776266									

The fitted equation is: $\hat{Y} = 89.12 - 9.01X$ and the estimated error variance is $s^2 = MSE = 50.78$, with corresponding standard deviation $s = 7.13$.

As tissue concentration of LSD increases by 1 unit, math scores tend to drop on average by 9.01 points.



Chapter 2 – Inferences in Regression Analysis

Rules Concerning Linear Functions of Random Variables

Let $Y_1,\ldots,Y_n$ be n random variables. Consider the function $\sum_{i=1}^{n} a_i Y_i$ where the coefficients $a_1,\ldots,a_n$ are constants. Then, we have:

$$E\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i E\{Y_i\} \qquad \sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{Y_i, Y_j\}$$

When $Y_1,\ldots,Y_n$ are independent (as in the model in Chapter 1), the variance of the linear combination simplifies to:

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 \sigma^2\{Y_i\}$$

When $Y_1,\ldots,Y_n$ are independent, the covariance of two linear functions $\sum_{i=1}^{n} a_i Y_i$ and $\sum_{i=1}^{n} c_i Y_i$ can be written as:

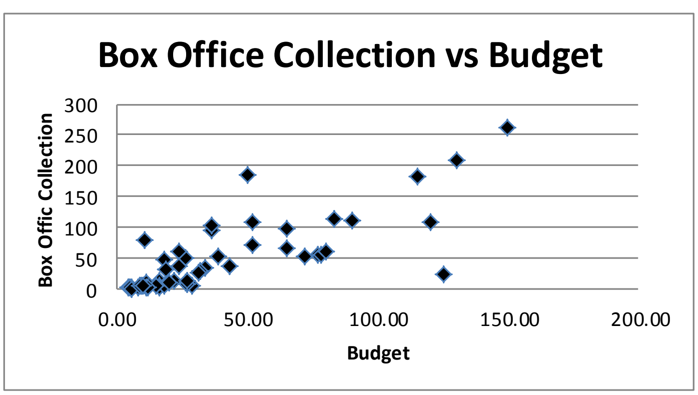
$$\sigma\left\{\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i,\right\} = \sum_{i=1}^{n} a_i c_i \sigma^2\{Y_i\}$$

We will use these rules to obtain the distribution of the estimators $b_0, b_1, \hat{Y} = b_0 + b_1 X$

Example: Bollywood Movie Budgets (X) and Box Office Grosses (Y)

Data: A sample of n = 55 Bollywood films released in 2013-2014. Data in crore, not certain of units.

http://www.bollymoviereviewz.com/2013/04/bollywood-box-office-collection-2013.html



X-bar	Y-Bar	SS_XX	SS_YY	SS_XY
39.04	46.88	72165.43	183601.1	90278.06

Inferences Concerning β_1

Recall that the least squares estimate of the slope parameter, b_1 , is a linear function of the observed responses $Y_1, \ldots, Y_n$:

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} Y_i = \sum_{i=1}^{n} k_i Y_i \qquad k_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{(X_i - \overline{X})}{SS_{XX}}$$

Note that $E\{Y_i\} = \beta_0 + \beta_1 X_i$, so that the expected value of b_1 is:

$$E\{b_1\} = \sum_{i=1}^{n} k_i E\{Y_i\} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{SS_{XX}} (\beta_0 + \beta_1 X_i) = \frac{1}{SS_{XX}} \left\{ \beta_0 \sum_{i=1}^{n} (X_i - \overline{X}) + \beta_1 \sum_{i=1}^{n} (X_i - \overline{X}) X_i \right\}$$

Note that $\sum_{i=1}^{n} (X_i - \overline{X}) = 0$ (why?), so that the first term in the brackets is 0, and that we can subtract $\beta_1 \overline{X} \sum_{i=1}^{n} (X_i - \overline{X}) = 0$ from the last term to get:

$$E\{b_1\} = \frac{1}{SS_{XX}} \left\{ \beta_1 \sum_{i=1}^{n} (X_i - \overline{X}) X_i - \beta_1 \sum_{i=1}^{n} (X_i - \overline{X}) \overline{X} \right\} = \frac{1}{SS_{XX}} \beta_1 \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{SS_{XX}} \beta_1 SS_{XX} = \beta_1$$

Thus, b_1 is an unbiased estimator of the parameter β_1.

Example: Bollywood Movie Data:

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{90278.06}{72165.43} = 1.2510$$

To obtain the variance of b_1, recall that $\sigma^2\{Y_i\} = \sigma^2$. Thus:

$$\sigma^2\{b_1\} = \sum_{i=1}^{n} k_i^2 \sigma^2\{Y_i\} = \sum_{i=1}^{n} \left[\frac{(X_i - \overline{X})}{SS_{XX}} \right]^2 \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{[SS_{XX}]^2} \sigma^2 = \frac{SS_{XX}}{[SS_{XX}]^2} \sigma^2 = \frac{\sigma^2}{SS_{XX}}$$

Note that the variance of b_1 decreases when we have larger sample sizes (as long as the added X levels are not placed at the sample mean $\overline{X}$). Since σ^2 is unknown in practice, and must be estimated from the data, we obtain the estimated variance of the estimator b_1 by replacing the unknown σ^2 with its unbiased estimate $s^2 = MSE$:

$$s^2\{b_1\} = \frac{s^2}{SS_{XX}} = \frac{MSE}{SS_{XX}}$$

with estimated standard error:

$$s\{b_1\} = \frac{s}{\sqrt{SS_{XX}}} = \frac{\sqrt{MSE}}{\sqrt{SS_{XX}}}$$

Example: Bollywood Movie Data:

$$SSE = SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}} = 183601.1 - \frac{90278.06^2}{72165.43} = 70664.4 \quad \Rightarrow \quad s^2 = MSE = \frac{SSE}{n-2} = \frac{70664.4}{55-2} = 1333.29$$

$$s^2\{b_1\} = \frac{s^2}{SS_{XX}} = \frac{1333.29}{72165.43} = 0.018475 \quad \Rightarrow \quad s\{b_1\} = \sqrt{0.018475} = 0.1359$$

Further, the sampling distribution of b_1 is normal, that is:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

Since, under the current model, b_1 is a linear function of independent, normal random variables $Y_1, \ldots, Y_n$. Making use of theory from mathematical statistics, we obtain the following result that allows us to make inferences concerning β_1:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$$

where $t(n\text{-}2)$ represents Student's t-distribution with n-2 degrees of freedom.

Confidence Interval for β_1

As a result of the fact that $\frac{b_1-\beta_1}{s\{b_1\}} \sim t(n-2)$, we obtain the following probability statement:

$$P\left\{t(\alpha/2;n-2) \le \frac{b_1-\beta_1}{s\{b_1\}} \le t(1-\alpha/2;n-2)\right\} = 1-\alpha$$

where $t(\alpha/2;n-2)$ is the $(\alpha/2)100^{th}$ percentile of the t-distribution with n-2 degrees of freedom. Note that since the t-distribution is symmetric around 0, we have that $t(\alpha/2;n-2) = -t(1-\alpha/2;n-2)$. We obtain the values corresponding to $t(1-\alpha/2;n-2)$ from tables or computer software, which is the value of that leaves an upper tail area of $\alpha/2$. The following algebra results in obtaining a $(1-\alpha)100\%$ confidence interval for β_1:

$$\begin{aligned}
&P\left\{t(\alpha/2;n-2) \le \frac{b_1-\beta_1}{s\{b_1\}} \le t(1-\alpha/2;n-2)\right\}\\
&= P\left\{-t(1-\alpha/2;n-2) \le \frac{b_1-\beta_1}{s\{b_1\}} \le t(1-\alpha/2;n-2)\right\}\\
&= P\left\{-t(1-\alpha/2;n-2)s\{b_1\} \le b_1-\beta_1 \le t(1-\alpha/2;n-2)s\{b_1\}\right\}\\
&= P\left\{-b_1-t(1-\alpha/2;n-2)s\{b_1\} \le -\beta_1 \le -b_1+t(1-\alpha/2;n-2)s\{b_1\}\right\}\\
&= P\left\{b_1+t(1-\alpha/2;n-2)s\{b_1\} \ge \beta_1 \ge b_1-t(1-\alpha/2;n-2)s\{b_1\}\right\}
\end{aligned}$$

This leads to the following rule for a $(1-\alpha)100\%$ confidence interval for β_1:

$$\boxed{b_1 \pm t(1-\alpha/2;n-2)s\{b_1\}}$$

Some statistical software packages print this out automatically (e.g. EXCEL and SPSS). Other packages simply print out estimates, standard errors, and t-statistics only, but have options to print them (e.g. R).

Example: Bollywood Movie Data:

$$t(.975;53) = 2.0057 \qquad b_1 = 1.2510 \qquad s\{b_1\} = 0.1359$$

$$95\% \text{ CI for } \beta_1: 1.2510 \pm 2.0057(0.1359) \equiv 1.2510 \pm 0.2726 \equiv (0.9784, 1.5236)$$

Tests Concerning β_1

We can also make use of the of the fact that $\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$ to test hypotheses concerning the slope parameter. As with means and proportions (and differences of means and proportions), we can conduct one-sided and two-sided tests, depending on whether a priori a specific directional belief is held regarding the slope. More often than not (but not necessarily), the null value for β_1 is 0 (the mean of Y is independent of X) and the alternative is that β_1 is positive (1-sided), negative (1-sided), or different from 0 (2-sided). The alternative hypothesis must be selected before observing the data. Default t-tests produced by computer software packages are two-sided tests that $\beta_1 = 0$.

2-sided tests

- Null Hypothesis: $H_0 : \beta_1 = \beta_{10}$
- Alternative (Research Hypothesis): $H_A : \beta_1 \neq \beta_{10}$
- Test Statistic: $t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$
- Decision Rule: Conclude H_A if $|t^*| \geq t(1-\alpha/2; n-2)$, otherwise conclude H_0
- P-value: $2P(t(n-2) > |t^*|)$

All statistical software packages (to my knowledge) will print out the test statistic and P-value corresponding to a 2-sided test with β_{10}=0.

1-sided tests (Upper Tail)

- Null Hypothesis: $H_0 : \beta_1 \leq \beta_{10}$
- Alternative (Research Hypothesis): $H_A : \beta_1 > \beta_{10}$
- Test Statistic: $t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$
- Decision Rule: Conclude H_A if $t^* \geq t(1-\alpha; n-2)$, otherwise conclude H_0
- P-value: $P(t(n-2) > t^*)$

A test for positive association between Y and X (H_A:β_1>0) can be obtained from standard statisical software by first checking that b_1 (and thus t^*) is positive, and cutting the printed P-value in half.

1-sided tests (Lower Tail)

- Null Hypothesis: $H_0 : \beta_1 \geq \beta_{10}$
- Alternative (Research Hypothesis): $H_A : \beta_1 < \beta_{10}$
- Test Statistic: $t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$
- Decision Rule: Conclude H_A if $t^* \leq -t(1-\alpha; n-2)$, otherwise conclude H_0
- P-value: $P(t(n-2) < t^*)$

A test for negative association between Y and X (H_A:β_1<0) can be obtained from standard statistical software by first checking that b_1 (and thus t^*) is negative, and cutting the printed P-value in half.

Example: Bollywood Movie Data:

Question 1: Is there any association between Box Office Collection and Budget?
Question 2: Does increasing Budget by 1 unit lead to an increase in average Box Office Collection by > 1 unit?

$$\text{Q1:}\ H_0^1:\beta_1=0 \quad H_A^1:\beta_1\neq 0 \qquad \text{Q2:}\ H_0^2:\beta_1\leq 0 \quad H_A^2:\beta_1>0$$

$$\text{TS1:}\ t_1^*=\frac{1.2510-0}{0.1359}=9.2035 \qquad t(.975;53)=2.0057 \quad \text{Decision?}$$

$$P-\text{value:}\quad 2P\left(t(53)\geq|9.2035|\right)\approx 0$$

$$\text{TS2:}\ t_2^*=\frac{1.2510-1}{0.1359}=1.8469 \qquad t(.95;53)=1.6741 \quad \text{Decision?}$$

$$P-\text{value:}\quad P\left(t(53)\geq 1.8469\right)=.0352$$

Inferences Concerning β_0

Recall that the least squares estimate of the intercept parameter, b_0 , is a linear function of the observed responses $Y_1,\ldots,Y_n$:

$$b_0=\overline{Y}-b_1\overline{X}=\sum_{i=1}^{n}\left[\frac{1}{n}+\frac{(X_i-\overline{X})\overline{X}}{SS_{XX}}\right]Y_i=\sum_{i=1}^{n}l_iY_i$$

Recalling that $E\{Y_i\}=\beta_0+\beta_1X_i$:

$$E\{b_0\}=\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{(X_i-\overline{X})\overline{X}}{SS_{XX}}\right](\beta_0+\beta_1X_i) \quad = \quad \beta_0\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{(X_i-\overline{X})\overline{X}}{SS_{XX}}\right]+\beta_1\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{(X_i-\overline{X})\overline{X}}{SS_{XX}}\right]X_i$$

$$=\beta_0(1-0)+\beta_1\left[\frac{1}{n}\sum_{i=1}^{n}X_i-\overline{X}\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{SS_{XX}}\right]=\beta_0+\beta_1(\overline{X}-\overline{X}(1))=\beta_0$$

Thus, b_0 is an unbiased estimator or the parameter β_0.

Example: Bollywood Movie Data: $\boxed{b_0=\overline{Y}-b_1\overline{X}=46.88-1.2510(39.04)=-1.9549}$

Below, we obtain the variance of the estimator of b_0.

$$\sigma^2\{b_0\} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{(X_i - \overline{X})\overline{X}}{SS_{XX}}\right]^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n}\left[\frac{1}{n^2} + \frac{\overline{X}^2 (X_i - \overline{X})^2}{\left(SS_{XX}\right)^2} - \frac{2\overline{X}(X_i - \overline{X})}{nSS_{XX}}\right]$$

$$= \sigma^2 \left[\frac{n}{n^2} + \frac{\overline{X}^2}{\left(SS_{XX}\right)^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 - \frac{2\overline{X}}{nSS_{XX}}\sum_{i=1}^{n}(X_i - \overline{X})\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right]$$

Note that the variance will decrease as the sample size increases, as long as X values are not all placed at the mean (which would not allow the regression to be fit). Further, the sampling distribution is normal under the assumptions of the model. The estimated standard error of b_0 replaces σ^2 with its unbiased estimate $s^2=MSE$ and taking the square root of the variance.

$$s\{b_0\} = s\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}} = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right]}$$

Example: Bollywood Movie Data:

$$MSE = s^2 = 1333.29 \quad n = 55 \quad \overline{X} = 39.04 \quad SS_{XX} = 72165.43$$

$$s\left\{b_0\right\} = \sqrt{1333.29\left[\frac{1}{55} + \frac{39.04^2}{72165.43}\right]} = \sqrt{52.40} = 7.24$$

Note that $\frac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2)$, allowing for inferences concerning the intercept parameter β_0 when it is meaningful, namely when X=0 is within the range of observed data.

Confidence Interval for β_0

$$b_0 \pm t(1-\alpha/2; n-2)s\{b_0\}$$

Example: Bollywood Movie Data:

Although no movies have a budget of X=0, a 95% CI for β_0 would be computed as follows:

$$-1.9549 \pm 2.0057(7.2385) \equiv -1.9549 \pm 14.5185 \equiv (-16.47, 12.56)$$

It is also useful to obtain the covariance of b_0 and b_1, as they are only independent under very rare circumstances:

$$\sigma\{b_0,b_1\}=\sigma\left\{\sum_{i=1}^{n} l_iY_i,\sum_{i=1}^{n} k_iY_i\right\}=\sum_{i=1}^{n} l_ik_i\sigma^2\{Y_i\} \quad = \quad \sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{SS_{XX}}\right]\frac{(X_i-\overline{X})}{SS_{XX}}\sigma^2$$

$$=\frac{\sigma^2}{nSS_{XX}}\sum_{i=1}^{n}(X_i-\overline{X})-\frac{\sigma^2\overline{X}}{\left(SS_{XX}\right)^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 \quad = \quad 0-\frac{\sigma^2\overline{X}}{SS_{XX}} \quad = \quad -\frac{\sigma^2\overline{X}}{SS_{XX}}$$

In practice, $\overline{X}$ is usually positive, so that the intercept and slope estimators are usually negatively correlated. We will use the result shortly.

Considerations on Making Inferences Concerning β_0 and β_1

Normality of Error Terms

If the data are approximately normal, simulation results have shown that using the *t*-distribution will provide approximately correct significance levels and confidence coefficients for tests and confidence intervals, respectively. Even if the distribution of the errors (and thus Y) is far from normal, in large samples the sampling distributions of b_0 and b_1 have sampling distributions that are approximately normal as results of central limit theorems. This is sometimes referred to as *asymptotic normality*.

Interpretations of Confidence Coefficients and Error Probabilities

Since X levels are treated as fixed constants, these refer to the case where we repeated the experiment many times at the current set of X levels in this data set. In this sense, it's easier to interpret these terms in controlled experiments where the experimenter has set the levels of X (such as time and temperature in a laboratory type setting) as opposed to observational studies, where nature determines the X levels, and we may not be able to reproduce the same conditions repeatedly. This will be covered later.

Spacing of *X* Levels

The variances of b_0 and b_1 (for given n and σ^2) decrease as the X levels are more spread out, since their variances are inversely related to $SS_{XX}=\sum_{i=1}^{n}(X_i-\overline{X})^2$. However, there are reasons to choose a diverse range of X levels for assessing model fit. This is covered in Chapter 4.

Power of Tests

The **power** of a statistical test refers to the probability that we reject the null hypothesis. Note that when the null hypothesis is true, the power is simply the probability of a Type I error (α). When the null hypothesis is false, the power is the probability that we correctly reject the null hypothesis, which is 1 minus the probability of a Type II error (π=1-β), where π denotes the power of the test and β is the probability of a Type II error (failing to reject the null hypothesis when the alternative hypothesis is true). The following procedure can be used to obtain the power of the test concerning the slope parameter with a 2-sided alternative.

1) Write out null and alternative hypotheses: $H_0 : \beta_1 = \beta_{10} \qquad H_A : \beta_1 \neq \beta_{10}$
2) Obtain the non-centrality measure, the standardized distance between the true value of β_1 and the value under the null hypothesis (β_{10}): $\delta = \dfrac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}}$
3) Choose the probability of a Type I error (α=0.05 or α=0.01)
4) Determine the degrees of freedom for error: $df = n$-2
5) Using R, we can obtain the power as: Power = 1-pf(qf(1-α,1,n-2),1,n-2,δ^2)

Note that the power increases as the non-centrality measure increases for a given degrees of freedom, and as the degrees of freedom increases for a given non-centrality measure.

Confidence Interval for E{Y_h}=β_0+$\beta_1$$X_h$

When we wish to estimate the mean at a hypothetical X value (within the range of observed X values), we can use the fitted equation at that value of X=X_h as a **point estimate**, but we have to include the uncertainty in the regression estimators to construct a confidence interval for the mean.

Parameter: $E\{Y_h\} = \beta_0 + \beta_1 X_h$

Estimator: $\hat{Y}_h = b_0 + b_1 X_h$

We can obtain the variance of the estimator (as a function of X=X_h) as follows:

$$\boxed{\begin{aligned}&\sigma^2\left\{\hat{Y}_h\right\} = \sigma^2\{b_0 + b_1 X_h\} = \sigma^2\{b_0\} + X_h^2\sigma^2\{b_1\} + 2X_h\sigma\{b_0, b_1\}\\ &= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right] + X_h^2\frac{\sigma^2}{SS_{XX}} + 2X_h\left[-\frac{\sigma^2\overline{X}}{SS_{XX}}\right] \quad = \quad \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]\end{aligned}}$$

Estimated standard error of estimator: $s\{\hat{Y}_h\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$

Example: Bollywood Movie Data:

Suppose we are interested in **mean** Box Office Collection of **all possible** movies with budgets of $X_h = 20$

$$\boxed{\begin{array}{l} X_h = 20 \quad b_0 = -1.9549 \quad b_1 = 1.2510 \quad \hat{Y}_h = -1.9549 + 1.2510(20) = 23.07 \\ MSE = s^2 = 1333.29 \quad n = 55 \quad X_h = 20 \quad \overline{X} = 39.04 \quad SS_{XX} = 72165.43 \\ s\left\{\hat{Y}_h\right\} = \sqrt{1333.29\left[\frac{1}{55} + \frac{(20-39.04)^2}{72165.43}\right]} = \sqrt{1333.29(0.02321)} = \sqrt{30.94} = 5.56 \end{array}}$$

$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \sim t(n-2)$ which can be used to construct confidence intervals for the mean response at specific X levels, and tests concerning the mean (tests are rarely conducted).

(1-α)100% Confidence Interval for $E\{Y_h\}$:

$$\boxed{\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}}$$

Example: Bollywood Movie Data:

$$\boxed{23.07 \pm 2.0057(5.56) \equiv 23.07 \pm 11.15 \equiv (11.92, 34.22)}$$

Predicting a Future Observation When X is Known

If β_0, β_1, σ were known, we'd know that the distribution of responses when X=X_h is normal with mean $\beta_0 + \beta_1 X_h$ and standard deviation σ. Thus, making use of the normal distribution (and equivalently, the empirical rule) we know that if we took a sample item from this distribution, it is very likely that the value will fall within 2 standard deviations of the mean. That is, we would know that the probability that the sampled item lies within the range $(\beta_0 + \beta_1 X_h - 2\sigma, \beta_0 + \beta_1 X_h + 2\sigma)$ is approximately 0.95.

In practice, we don't know the mean $\beta_0 + \beta_1 X_h$ or the standard deviation σ. However, we have just constructed a (1-α)100% Confidence Interval for $E\{Y_h\}$, and we have an estimate of σ (s). Intuitively, we can approximately use the logic of the previous paragraph (with the estimate of σ) across the range of believable values for the mean. Then our prediction interval spans the lower tail of the normal curve centered at the lower bound for the mean to the upper tail of the normal curve centered at the upper bound for the mean.

The prediction error for the new observation is the difference between the observed value and its predicted value: $Y_h - \hat{Y}_h$. Since the data are assumed to be independent, the new (future) value is independent of its predicted value, since it wasn't used in the regression analysis. The variance of the prediction error can be obtained as follows:

$$\sigma^2\{pred\} = \sigma^2\{Y_h - \hat{Y}_h\} = \sigma^2\{Y_h\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$
$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

and an unbiased estimator is:

$$s^2\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$$

Example: Bollywood Movie Data:

Suppose interested in **predicting** Box Office Collection of **a single new** movie with a budget of $X_h = 20$

$$X_h = 20 \quad b_0 = -1.9549 \quad b_1 = 1.2510 \quad \hat{Y}_h = -1.9549 + 1.2510(20) = 23.07$$
$$MSE = s^2 = 1333.29 \quad n = 55 \quad X_h = 20 \quad \overline{X} = 39.04 \quad SS_{XX} = 72165.43$$
$$s\{pred\} = \sqrt{1333.29\left[1 + \frac{1}{55} + \frac{(20 - 39.04)^2}{72165.43}\right]} = \sqrt{1333.29(1.02321)} = \sqrt{1364.24} = 36.94$$

(1-α)100% Prediction Interval for New Observation When X=X_h

$$\hat{Y}_h \pm t(\alpha/2; n-2)\sqrt{MSE\left[1+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]}$$

Example: Bollywood Movie Data:

$$23.07 \pm 2.0057(36.94) \equiv 23.07 \pm 74.08 \equiv (-51.01, 97.15) \equiv (0, 97.15)$$

Note: Unlike a Confidence Interval for a mean, which has a standard error that gets smaller, as the sample size increases, the Prediction Interval for a single observation cannot be smaller than *s*, the residual standard deviation. When that is large, prediction intervals will be wide, and often of little use.

It is a simple extension to obtain a prediction for the mean of *m* new observations when X=X_h. The sample mean of *m* observations is $\frac{\sigma^2}{m}$ and we get the following variance for the error in the prediction mean.

$$s^2\{predmean\} = MSE\left[\frac{1}{m}+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]$$

(1-α)100% Prediction Interval for the Mean of *m* New Observations When X=X_h

$$\hat{Y}_h \pm t(\alpha/2; n-2)\sqrt{MSE\left[\frac{1}{m}+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]}$$

(1-α)100% Confidence Band for the Entire Regression Line (Working-Hotelling Method)

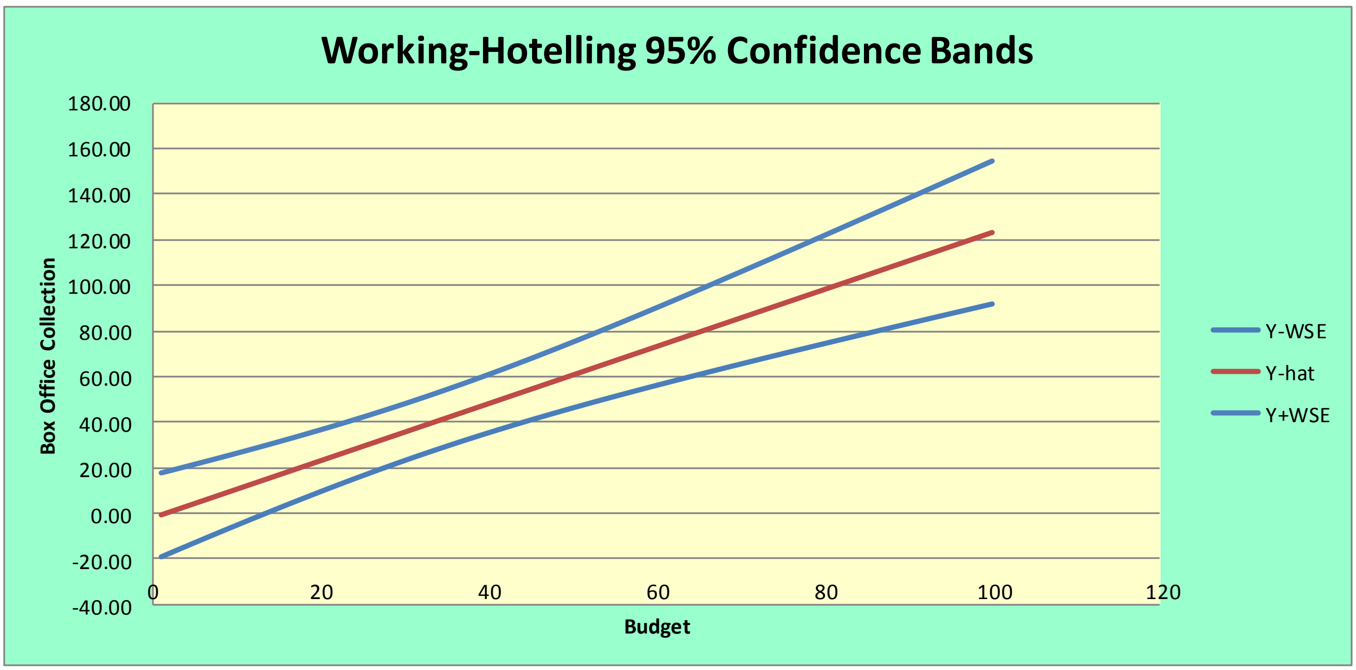
$$\hat{Y}_h \pm Ws\{\hat{Y}_h\} \qquad W = \sqrt{2F(1-\alpha; 2, n-2)}$$

Example: Bollywood Movie Data:

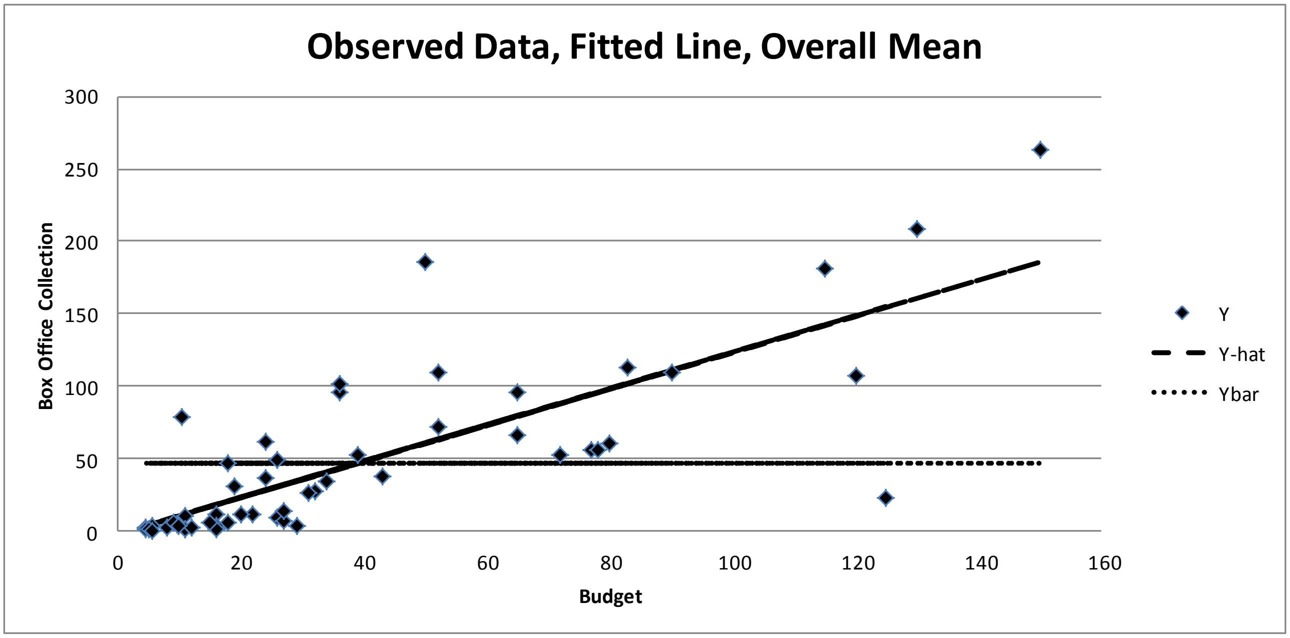
$$W = \sqrt{2F\left(0.95; 2, 55-2\right)} = \sqrt{2(3.1716)} = 2.5186$$

Selected values of X_h , estimates, standard errors, and half-widths for confidence band:

X_h	Y-hat	SE{Y-hat}	W*SE{Yh}
5	4.30	6.82	17.18
20	23.06	5.38	13.55
40	48.08	5.07	12.78
60	73.10	6.76	17.02
80	98.12	9.42	23.73
100	123.14	12.45	31.35



Analysis of Variance Approach to Regression



Consider the total deviations of the observed responses from the mean: $Y_i - \bar{Y}$. When these terms are all squared and summed up, this is referred to as the **total sum of squares (SSTO)**.

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

In the plot, these are the vertical distance of the points to the purple line just below 50. The more spread out the observed data are, the larger SSTO will be.

Now consider the deviation of the observed responses from their fitted values based on the regression model: $Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) = e_i$. When these terms are squared and summed up, this is referred to as the **error sum of squares (SSE)**. We've already encountered this quantity and used it to estimate the error variance.

$$\boxed{SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

When the observed responses fall close to the regression line, SSE will be small. When the data are not near the line, SSE will be large.

Finally, there is a third quantity, representing the deviations of the predicted values from the mean. Then these deviations are squared and summed up, this is referred to as the **regression sum of squares** **(SSR)**.

$$\boxed{SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}$$

The error and regression sums of squares sum to the total sum of squares: $SSTO = SSR + SSE$ which can be seen as follows:

$$Y_i - \overline{Y} = Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}) \quad \Rightarrow$$
$$(Y_i - \overline{Y})^2 = [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})]^2 = (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \overline{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) \quad \Rightarrow$$
$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \overline{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y})\right] =$$
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}e_i(b_0 + b_1X_i - \overline{Y}) =$$
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\left[b_0\sum_{i=1}^{n}e_i + b_1\sum_{i=1}^{n}e_iX_i - \overline{Y}\sum_{i=1}^{n}e_i\right] = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2(0) =$$
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = SSE + SSR$$

The last term was 0 since $\sum e_i = \sum e_i X_i = 0$,

Each sum of squares has associated with **degrees of freedom**. The total degrees of freedom is $df_T = n$-1. The error degrees of freedom is $df_E = n$-2. The regression degrees of freedom is $df_R = 1$. Note that the error and regression degrees of freedom sum to the total degrees of freedom: $n-1 = 1 + (n-2)$.

Mean squares are the sums of squares divided by their degrees of freedom:

$$MSR = \frac{SSR}{1} \qquad MSE = \frac{SSE}{n-2}$$

Note that *MSE* was our estimate of the error variance, and that we don't compute a total mean square. It can be shown that the expected values of the mean squares are:

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \beta_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Note that these expected mean squares are the same if and only if β_1=0.

The Analysis of Variance is reported in tabular form:

Source	df	SS	MS	F
Regression	1	SSR	MSR=SSR/1	F=MSR/MSE
Error	*n*-2	SSE	MSE=SSE/(*n*-2)	
C Total	*n*-1	SSTO		

Example: Bollywood Movie Data:

$$\text{Total:} \quad SSTO = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = SS_{YY} = 183601.1 \quad df_{TO} = 55 - 1 = 54$$

$$\text{Error (Residual):} \quad SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = SS_{YY} - \frac{SS_{XY}^2}{SS_{XX}} = 183601.1 - \frac{(90278.6)^2}{72165.43} = 70664.4 \quad df_E = 55 - 2 = 53$$

$$\text{Regression:} \ SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \frac{SS_{XY}^2}{SS_{XX}} = \frac{(90278.6)^2}{72165.43} = 112936.7 \quad df_R = 1$$

ANOVA Table:

ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	112936.7	112936.7	84.70529
Residual	53	70664.39	1333.29	
Total	54	183601.1		

F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

As a result of **Cochran's Theorem (stated on page 76 of text book),** we have a test of whether the dependent variable *Y* is linearly related to the predictor variable *X*. This is a very specific case of the *t*-test described previously. Its full utility will be seen when we consider multiple predictors. The test proceeds as follows:

- Null hypothesis: $H_0 : \beta_1 = 0$
- Alternative (Research) Hypothesis: $H_A : \beta_1 \neq 0$
- Test Statistic: $TS : F^* = \dfrac{MSR}{MSE}$
- Rejection Region: $RR : F^* \geq F(1-\alpha;1,n-2)$
- *P*-value: $P\{F(1,n-2) \geq F^*\}$

Critical values of the *F*-distribution (indexed by numerator and denominator degrees' of freedom) are given in **Table B.4, pages 1340-1345**, and on class website, and can be obtained simply in EXCEL or R (see Introduction). P-values must be obtained in EXCEL or R.

Note that this is a very specific version of the *t*-test regarding the slope parameter, specifically a 2-sided test of whether the slope is 0. Mathematically, the tests are identical:

$$t^* = \frac{b_1 - 0}{s\{b_1\}} = \frac{\dfrac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}}{\sqrt{\dfrac{MSE}{\sum (X_i - \overline{X})^2}}} = \frac{\dfrac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}}}{\sqrt{MSE}} = \frac{\dfrac{SS_{XY}}{\sqrt{SS_{XX}}}}{\sqrt{MSE}} = \frac{\sqrt{\dfrac{SS_{XY}^2 / SS_{XX}}{1}}}{\sqrt{MSE}} = \frac{\sqrt{MSR}}{\sqrt{MSE}}$$

$$\Rightarrow \quad (t^*)^2 = \frac{MSR}{MSE} = F^*$$

Further, the critical values are equivalent: $(t(1-\alpha/2;n-2))^2 = F(1-\alpha;1,n-2)$,

check this from the two tables. Thus, the tests are equivalent.

Example: Bollywood Movie Data:

$$H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$$
$$\text{Test Statistic: } F^* = \frac{MSR}{MSE} = \frac{112936.7}{1333.29} = 84.71 \qquad F\left(0.95;1,55-2\right) = 4.023$$
$$P-\text{value: } P\left(F(1,53) \geq 84.71\right) = .0000$$

Confirm the *t*-statistic, when squared, gives the *F*-statistic, and that the critical *t*-value for the 2-sided *t*-test is the same as the critical *F*-value.

General Linear Test Approach

This is a very general method of testing hypotheses concerning regression models. We first consider the the simple linear regression model, and testing whether *Y* is linearly associated with *X*. We wish to test $H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$.

Full Model

This is the model specified under the alternative hypothesis, also referred to as the unrestricted model. Under simple linear regression with normal errors, we have:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Using least squares (and maximum likelihood) to estimate the model parameters and the fitted values ($\hat{Y}_i = b_0 + b_1 X_i$), we obtain the error sum of squares for the full model:

$$SSE(F) = \sum (Y_i - (b_0 + b_1 X_i))^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE$$

Reduced Model

This the model specified by the null hypothesis, also referred to as the restricted model. Under simple linear regression with normal errors, we have:

$$Y_i = \beta_0 + 0X_i + \varepsilon_i = \beta_0 + \varepsilon_i$$

Using least squares (and maximum likelihood) to estimate the model parameter, we obtain $\overline{Y}$ as the estimate of β_0, and have $b_0 = \overline{Y}$ as the fitted value for each observation. We then obtain the following error sum of squares under the reduced model:

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \overline{Y})^2 = SSTO$$

Test Statistic

The error sum of squares for the full model will always be less than or equal to the error sum of squares for reduced model, by definition of least squares. The test statistic will be:

$$\boxed{F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}}$$

where df_R, df_F are the error degrees of freedom for the full and reduced models. We will use this method throughout course.

For the simple linear regression model, we obtain the following quantities:

$$\boxed{SSE(F) = SSE \qquad df_F = n-2 \qquad SSE(R) = SSTO \qquad df_R = n-1}$$

thus the *F*-Statistic for the General Linear Test can be written:

$$\boxed{F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{SSTO - SSE}{(n-1)-(n-2)}}{\frac{SSE}{n-2}} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE}}$$

Thus, for this particular null hypothesis, the general linear test "generalizes" to the *F*-test.

<u>Example: Bollywood Movie Data:</u>

Suppose we wish to test whether on average, Box office collection is equal to the movie's budget.

$$E\{Y\} = X \quad \Rightarrow \quad H_0: \beta_0 = 0, \quad \beta_1 = 1 \quad \Rightarrow \quad SSE(R) = \sum_{i=1}^{n}(Y_i - X_i)^2 = 78593.4 \quad df_R = n = 55$$

$$SSE(F) = 70664.4 \quad df_F = 55 - 2 = 53$$

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{78593.4 - 70664.4}{55-53}}{\frac{70664.4}{53}} = \frac{3964.5}{1333.3} = 2.973 \qquad F(0.95; 2, 53) = 3.172 \quad P\text{-value} = 0.0597$$

<u>Descriptive Measures of Association</u>

Along with the slope, *Y*-intercept, and error variance; several other measures are often reported.

Coefficient of Determination (r^2)

The coefficient of determination measures the proportion of the variation in *Y* that is "explained" by the regression on *X*. It is computed as the regression sum of squares divided by the total (corrected) sum of squares. Values near 0 imply that the regression model has done little to "explain" variation in *Y*, while values near 1 imply that the model has "explained" a large portion of the variation in *Y*. If all the data fall exactly on the fitted line, r^2=1. The coefficient of determination will lie beween 0 and 1.

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad 0 \le r^2 \le 1$$

Coefficient of Correlation (*r*)

The coefficient of correlation is a measure of the strength of the linear association between *Y* and *X*. It will always be the same sign as the slope estimate (b_1), but it has several advantages:

- In some applications, we cannot identify a clear dependent and independent variable, we just wish to determine how two variables vary together in a population (peoples heights and weights, closing stock prices of two firms, etc). Unlike the slope estimate, the coefficient of correlation does not depend on which variable is labeled as *Y*, and which is labeled as *X*.
- The slope estimate depends on the units of *X* and *Y*, while the correlation coefficient does not.
- The slope estimate has no bound on its range of potential values. The correlation coefficient is bounded by –1 and +1, with higher values (in absolute value) implying stronger linear association (it is not useful in measuring nonlinear association which may exist, however).

$$r = \mathrm{sgn}(b_1)\sqrt{r^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})(Y_i - \overline{Y})}} = \frac{s_x}{s_y} b_1 \qquad -1 \le r \le 1$$

where $\mathrm{sgn}(b_1)$ is the sign (positive or negative) of b_1, and s_x, s_y are the sample standard deviations of *X* and *Y*, respectively.

<u>Example: Bollywood Movie Data:</u>

$$r^2 = \frac{SSR}{SSTO} = \frac{112936.7}{183601.1} = 0.6151$$

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} SS_{YY}}} = \frac{90278.06}{\sqrt{72165.43(183601.1)}} = .7843$$

Approximately 61.5% of the variation in box-office collection is “explained” by the film’s budget.

Tests Concerning the Population Correlation ρ

Parameter: $\rho_{12} = \dfrac{\sigma\{Y_1, Y_2\}}{\sigma\{Y_1\}\sigma\{Y_2\}} = \dfrac{\sigma_{12}}{\sigma_1\sigma_2}$

Point (maximum likelihood) Estimator (aka Pearson product-moment correlation coefficient):

$$r_{12} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} \qquad -1 \le r_{12} \le 1$$

Testing $H_0: \rho_{12} = 0$ vs $H_A: \rho_{12} \ne 0$:

Test Statistic: $t^* = \dfrac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}}$

Reject H_0 if $\left|t^*\right| \ge t\left(1-(\alpha/2); n-2\right)$

For 1-sided tests:

$H_A: \rho_{12} > 0$: Reject H_0 if $t^* \ge t\left(1-\alpha; n-2\right)$

$H_A: \rho_{12} < 0$: Reject H_0 if $t^* \le -t\left(1-\alpha; n-2\right)$

This test is mathematically equivalent to t-test for $H_0: \beta_1 = 0$

Example: Bollywood Movie Data:

$H_0: \rho_{12} = 0$ vs $H_A: \rho_{12} \ne 0$:

Test Statistic: $t^* = \dfrac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = \dfrac{.7843\sqrt{55-2}}{\sqrt{1-.6151}} = 9.204$

Reject H_0 if $\left|t^*\right| \ge t\left(1-(\alpha/2); n-2\right) = t\left(.975, 55-2\right) = 2.0057$

(1-α)100% Confidence Inteval for ρ

Problem: When $\rho_{12} \ne 0$, sampling distribution of r_{12} is messy

Fisher's z transformation: $z' = \dfrac{1}{2}\ln\left(\dfrac{1+r_{12}}{1-r_{12}}\right)$

For large n (typically at least 25): $z' \overset{approx}{\sim} N\left(\zeta, \dfrac{1}{n-3}\right) \qquad \zeta = \dfrac{1}{2}\ln\left(\dfrac{1+\rho_{12}}{1-\rho_{12}}\right)$

Compute an approximate $(1-\alpha)100\%$ CI for ζ and transform back for ρ:

$(1-\alpha)100\%$ CI for ζ: $\quad z' \pm z\left(1-(\alpha/2)\right)\sqrt{\dfrac{1}{n-3}}$

After computing CI for ζ, use identity $\rho_{12} = \dfrac{e^{2\xi}-1}{e^{2\xi}+1}$

Example: Bollywood Movie Data:

Fisher's z transformation: $z' = \frac{1}{2}\ln\left(\frac{1+r_{12}}{1-r_{12}}\right) = \frac{1}{2}\ln\left(\frac{1+0.7843}{1-0.7843}\right) = 1.0564$

$(1-\alpha)100\%$ CI for ζ: $\quad z' \pm z\left(1-(\alpha/2)\right)\sqrt{\frac{1}{n-3}} \quad \equiv$

$1.0564 \pm 1.96\sqrt{\frac{1}{55-2}} \quad \equiv \quad 1.0564 \pm 0.2718 \quad \equiv \quad (0.7846, 1.3282)$

$\rho_{12,LB} = \frac{e^{2\xi_{LB}}-1}{e^{2\xi_{LB}}+1} = \frac{e^{2(0.7846)}-1}{e^{2(0.7846)}+1} = 0.6554 \qquad \rho_{12,UB} = \frac{e^{2\xi_{UB}}-1}{e^{2\xi_{UB}}+1} = \frac{e^{2(1.3282)}-1}{e^{2(1.3282)}+1} = 0.8688$

$\Rightarrow$ 95% CI for $\rho \equiv (0.6554, 0.8688)$

Issues in Applying Regression Analysis

- When using regression to predict the future, the assumption is that the conditions are the same in future as they are now. Clearly any future predictions of economic variables such as tourism made prior to September 11, 2001 would not be valid.

- Often when we predict in the future, we must also predict *X*, as well as *Y*, especially when we aren't controlling the levels of *X*. Prediction intervals using methods described previously will be too narrow (that is, they will overstate confidence levels).

- Inferences should be made only within the range of *X* values used in the regression analysis. We have no means of knowing whether a linear association continues outside the range observed. That is, we should not **extrapolate** outside the range of *X* levels observed in experiment.

- Even if we determine that *X* and *Y* are associated based on the *t*-test and/or *F*-test, we cannot conclude that changes in *X* **cause** changes in *Y*. Finding an association is only one step in demonstrating a causal relationship.

- When multiple tests and/or confidence intervals are being made, we must adjust our confidence levels. This is covered in Chapter 4.

- When X_i is a random variable, and not being controlled, all methods described thus far hold, as long as the X_i are independent, and their probability distribution does not depend on $\beta_0, \beta_1, \sigma^2$.