Math 1710 Class 22

Regression

Dr. Back

Oct. 19, 2009

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$-1 <= r <= 1$ ($= \pm 1$ only for perfect linear association)

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

r is unchanged if x and y are exchanged.

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

invariant under rescaling

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Curved association and r=0 are consistent!

Correlation Properties

r is strongly affected by outliers.

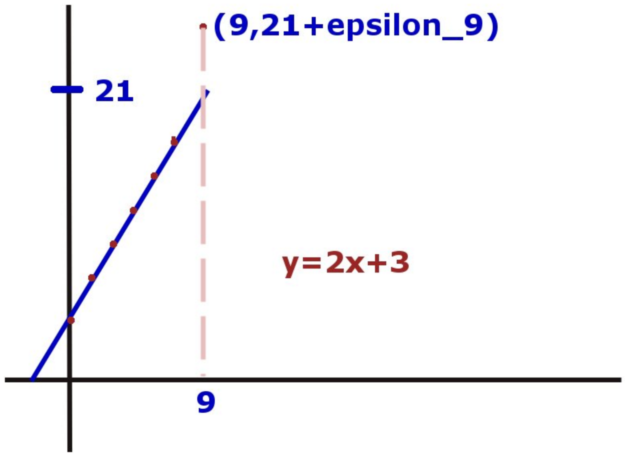
Correlation Properties

r is strongly affected by outliers. Consider 9 points (x=0, 1,. . . , 8 on the line $y = 2x + 3$ plus one outlier when $x = 9$.

x	y
0	3
1	5
2	7
...	...
8	19
9	$21 + \epsilon_9$

Correlation Properties

r is strongly affected by outliers.



Correlation Properties

r is strongly affected by outliers.

Correlation vs. size of ϵ_9.

ϵ_9	r
0	1.00
-10	.853
10	.944
-5	.968
-10	.853
-15	.663
-20	.455
-25	.275
5	.981
15	.904
20	.866
25	.834

Correlation Properties

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

samples from independent RV's $\Rightarrow \ r \sim 0$

Correlation Properties

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

X,Y indep std normal RV's ; set $Y^* = \rho X + \sqrt{1-\rho^2}\, Y$ Then (X, Y^*) will tend to generate data with $r \sim \rho$. (e.g. $\rho = .99$ $=> \frac{\sqrt{(1-\rho^2)}}{\rho} = .14$!)

What does Best Fitting Mean?

Given any point (x_i, y_i)

What does Best Fitting Mean?

Given any point (x_i, y_i)
and any line $y = c_0 + c_1 x$

What does Best Fitting Mean?

Given any point (x_i, y_i)
and any line $y = c_0 + c_1 x$
we can use the line to get a predicted value

$$\hat{y}_i = c_0 + c_1 x_i$$

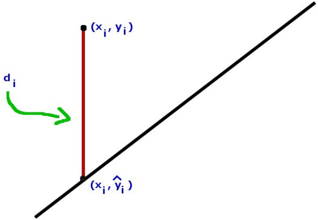
What does Best Fitting Mean?

Given any point (x_i, y_i)
and any line $y = c_0 + c_1 x$
we can use the line to get a predicted value

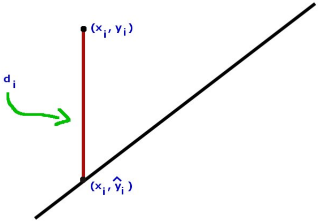
$$\hat{y}_i = c_0 + c_1 x_i$$

and define the residual d_i by

$$d_i = y_i - \hat{y}_i.$$

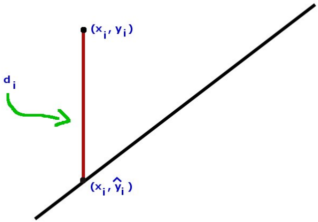


What does Best Fitting Mean?



Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

What does Best Fitting Mean?

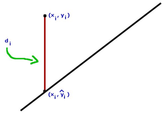


Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

Another try: Minimize Σd_i.

But cancellation would be a problem.

What does Best Fitting Mean?



Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

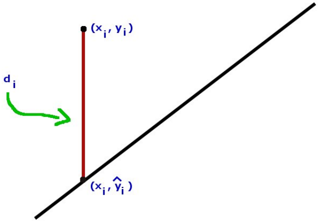
A Real Possibility: Minimize $\Sigma |d_i|$.

Not as nice a theory. (Abs value not differentiable)

Answer line can always be chosen to join two data points.

Sometimes used.

What does Best Fitting Mean?



Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

Answer: We'll show the best fitting line $\hat{y} = b_1 x + b_0$ is given by

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

What does Best Fitting Mean?

Best fitting means the line minimizing the sum Σd_i^2 of the squares of the vertical distances.

Answer: We'll show the best fitting line $\hat{y} = b_1 x + b_0$ is given by

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

i.e.: *Slope is r in standard deviation units.*

And

$$b_0 = \bar{y} - b_1\bar{x}$$

i.e.: *The point $(\bar{x}, \bar{y})$ lies on the line.*

What does Best Fitting Mean?

Strictly speaking, the word residual refers to this vertical distance d_i JUST in the case that the line is the “answer line”

$$\hat{y} = b_0 + b_1 x.$$

R^2

"The proportion of the variation in y explained by the regression of y on x is r^2."

R^2

"The proportion of the variation in y explained by the regression of y on x is r^2."

This actually means that for the 1-variable data sets $\{y_i\}$ and $\{\hat{y}_i\}$ (the predicted values) we have:

$$r^2 = \frac{Var(\hat{y}_i)}{Var(y_i)}$$

R^2

"The proportion of the variation in y explained by the regression of y on x is r^2."

This actually means that for the 1-variable data sets $\{y_i\}$ and $\{\hat{y}_i\}$ (the predicted values) we have:

$$r^2 = \frac{Var(\hat{y}_i)}{Var(y_i)}$$

More meaningful is the companion statement about residuals:

$$Var(d_i) = (1 - r^2) \cdot Var(y_i)$$

.

R^2

"The proportion of the variation in y explained by the regression of y on x is r^2."

This actually means that for the 1-variable data sets $\{y_i\}$ and $\{\hat{y}_i\}$ (the predicted values) we have:

$$r^2 = \frac{Var(\hat{y}_i)}{Var(y_i)}$$

More meaningful is the companion statement about residuals:

$$Var(d_i) = (1 - r^2) \cdot Var(y_i)$$

. Note

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) = \hat{y}_i + d_i$$

.

Proof of Formula for Line of Regression

Think of an arbitrary line as being given by:

$$(y - \bar{y}) = c(x - \bar{x}) + d$$

Proof of Formula for Line of Regression

$$(y - \bar{y}) = c(x - \bar{x}) + d$$

Look at the sum of the squares of the residuals of the n points:

$$\Sigma(y_i - c(x_i - \bar{x}) - (d + \bar{y}))^2 \quad =$$

$$\begin{aligned} &= \Sigma((y_i - \bar{y}) - c(x_i - \bar{x}) - d)^2 \\ &= \Sigma(y_i - \bar{y})^2 - 2c(x_i - \bar{x})(y_i - \bar{y}) + c^2(x_i - \bar{x})^2 \\ &\quad +\Sigma d^2 + 2cd(x_i - \bar{x}) - 2d(y_i - \bar{y}) \end{aligned}$$

Proof of Formula for Line of Regression

$$(y - \bar{y}) = c(x - \bar{x}) + d$$

$$\Sigma(y_i - \bar{y})^2 - 2c(x_i - \bar{x})(y_i - \bar{y}) + c^2(x_i - \bar{x})^2 + \Sigma d^2 + 2cd(x_i - \bar{x}) - 2d(y_i - \bar{y})$$

Use the definitions of s_x^2, s_y^2, and r. Also use the fact that $\Sigma(x_i - \bar{x}) = 0$ and $\Sigma(y_i - \bar{y}) = 0$. Then the above sum of squares becomes

$$(n-1)(s_y^2 - 2crs_x s_y + c^2 s_x^2) + nd^2 =$$

$$= (n-1)((cs_x - rs_y)^2 + (1 - r^2)s_y^2) + nd^2$$

Proof of Formula for Line of Regression

$$(y - \bar{y}) = c(x - \bar{x}) + d$$

$$(n-1)((cs_x - rs_y)^2 + (1 - r^2)s_y^2) + nd^2$$

Because squares are always non-negative, this sum is minimized when $d = 0$ and $cs_x - rs_y = 0$. In other words, the line minimizing the sum of the squares of the residuals is

$$(y - \bar{y}) = r\frac{s_y}{s_x}(x - \bar{x})$$

in agreement with our usual formulas.

Regression Conditions

Textbook:

- (Paired) Quantitative Variables $(x_i, y_i), \quad i = 1 \ldots n$
- No Outliers
- Straight Enough Condition

Regression Conditions

Textbook:

- (Paired) Quantitative Variables (x_i, y_i), $\quad i = 1 \dots n$
- No Outliers
- Straight Enough Condition

Regression Plots Should Appear to Have:

- No pattern in the residuals.
- Constant standard deviation in residuals.
- Residuals normally distributed with mean 0.

Regression Conditions

Regression Plots Should Appear to Have:

- No pattern in the residuals.
- Constant standard deviation in residuals.
- Residuals normally distributed with mean 0.

Ideal Linear Relationship:

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

with the errors ϵ_i independent of each other and all following $N(0, \sigma)$ for some constant standard deviation σ.

Regression Conditions

Regression Plots Should Appear to Have:

- No pattern in the residuals.
- Constant standard deviation in residuals.
- Residuals normally distributed with mean 0.

Ideal Linear Relationship:

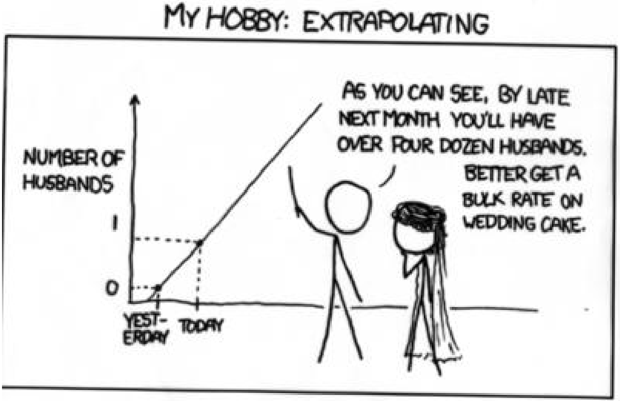
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

with the errors ϵ_i independent of each other and all following $N(0, \sigma)$ for some constant standard deviation σ.

There's no requirement that the x values be random.

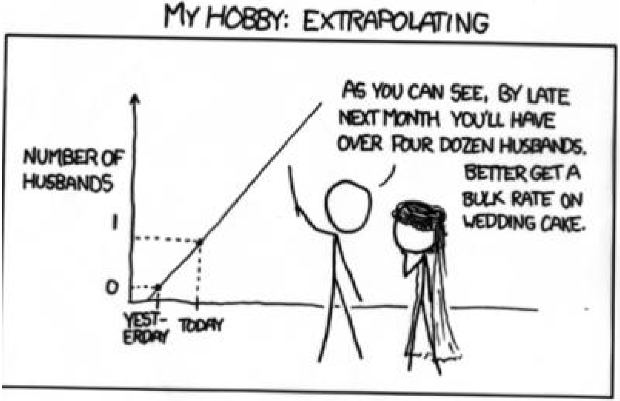
Regression Cautions

Extrapolation is Dangerous



Regression Cautions

Extrapolation is Dangerous



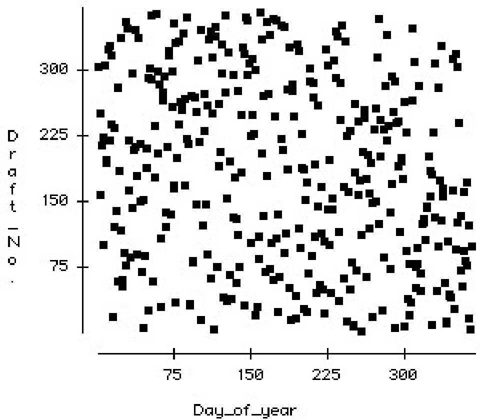
And watch out for lurking variables.

Was it Fair?

The first draft lottery during the Vietnam War:
366 balls labeled by dates.
Mixed up and pulled out in a “random” order.

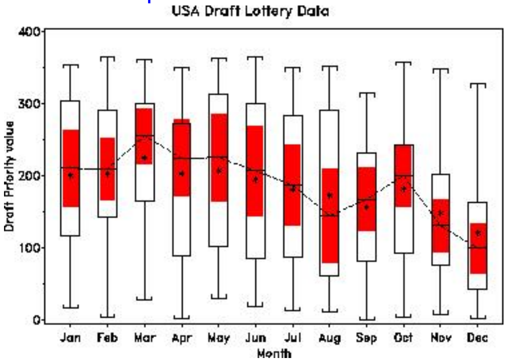
Was it Fair?

Scatterplot



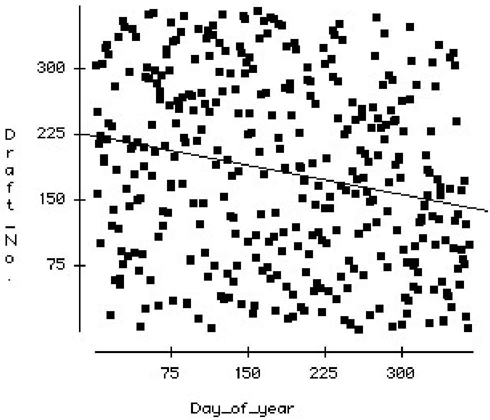
Was it Fair?

Boxplots for each month



Was it Fair?

Scatterplot with Regression Line



Was it Fair?

Correlation Display

Pearson Product-Moment Correlation

No Selector

	Draft__	Day_of_
Draft_No.	1.000	
Day_of_year	-0.226	1.000

Was it Fair?

Correlation Display

Pearson Product-Moment Correlation

No Selector

	Draft__	**Day_of_**
Draft_No.	1.000	
Day_of_year	-0.226	1.000

Around 1 in a thousand chance of a correlation coefficient this far from 0 if the lottery was fair.

Was it Fair?

Around 1 in a thousand chance of a correlation coefficient this far from 0 if the lottery was fair.

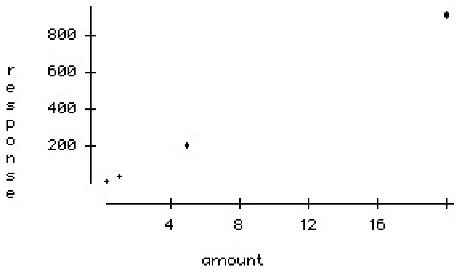
The balls were probably not mixed well enough.

Does R^2 near 1 Mean an Accurate Linear Model?

5 measurements each of 4 samples.
Amount of the substance in sample known in advance.
Response variable is the output reading from the gas chromatograph.

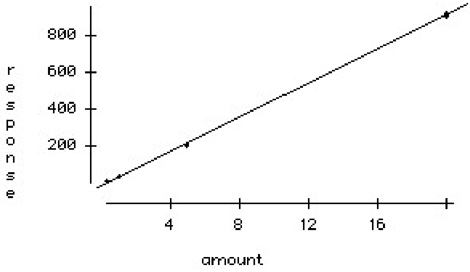
Does R^2 near 1 Mean an Accurate Linear Model?

Scatterplot



Does R^2 near 1 Mean an Accurate Linear Model?

Scatterplot with Regression Line



Does R^2 near 1 Mean an Accurate Linear Model?

Scatterplot with Regression Line

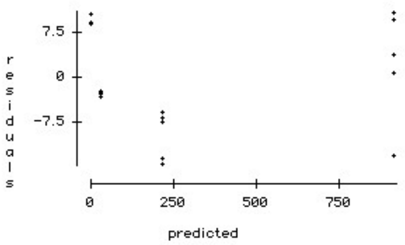
Dependent variable is: **response**
No Selector
21 total cases of which 1 is missing
R squared = 99.9%   R squared (adjusted) = 99.9%
s = 9.023 with 20 - 2 = 18 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	2.75907e6	1	2.75907e6	3.39e4
Residual	1465.53	18	81.4184	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	-14.4107	2.614	-5.51	≤ 0.0001
amount	46.6287	0.2533	184	≤ 0.0001

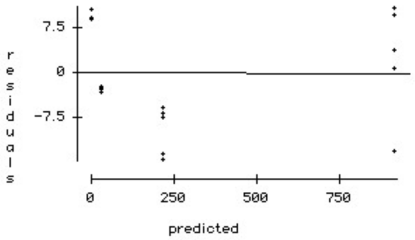
Does R^2 near 1 Mean an Accurate Linear Model?

Residual Plot



Does R^2 near 1 Mean an Accurate Linear Model?

Residual Plot with Horizontal Line



($\Sigma d_i = 0$ always.)

Does R^2 near 1 Mean an Accurate Linear Model?

Despite $r^2 = .999$, a linear model does not fully capture our situation here. Just plugging into the line of regression *would not* be the right way to make a prediction.

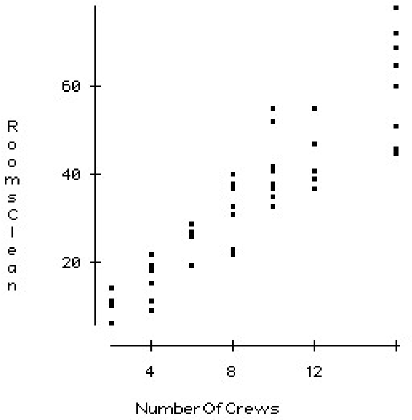
How Many Rooms Can x Crews Clean?

x crews working for a building contractor go out each night and clean y rooms.

Understand the relationship?

How Many Rooms Can x Crews Clean?

Scatterplot



NumCrews summary

Summary of Number Of Crews
No Selector
54 total cases of which 1 is missing

Percentile 25

Count	53
Mean	8.67925
Median	8
StdDev	4.80294
Range	14
IntQRange	8
Lower ith %tile	4
Upper ith %tile	12

How Many Rooms Can x Crews Clean?

RoomsCleaned Summary

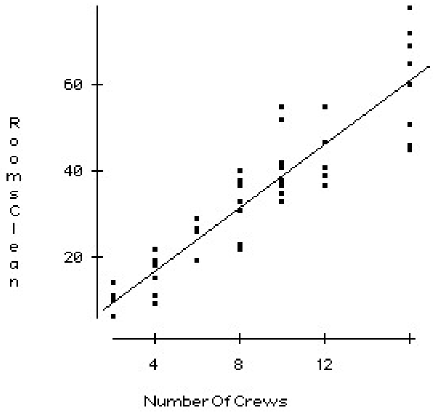
Summary of **RoomsClean**
No Selector
54 total cases of which 1 is missing

Percentile 25

Count	53
Mean	33.9057
Median	35
StdDev	19.2026
Range	72
IntQRange	27.5
Lower ith %tile	18.75
Upper ith %tile	46.25

How Many Rooms Can x Crews Clean?

Scatterplot with Regression Line



How Many Rooms Can x Crews Clean?

Regression Display

Dependent variable is: **RoomsClean**
No Selector
54 total cases of which 1 is missing
R squared = 85.7%   R squared (adjusted) = 85.4%
s = 7.336 with 53 - 2 = 51 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	16429.7	1	16429.7	305
Residual	2744.8	51	53.8195	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	1.7847	2.096	0.851	0.3986
NumberOfCr...	3.70089	0.2118	17.5	≤ 0.0001

How Many Rooms Can x Crews Clean?

Regression Display

Dependent variable is: **RoomsClean**
No Selector
54 total cases of which 1 is missing
R squared = 85.7%   R squared (adjusted) = 85.4%
s = 7.336 with 53 - 2 = 51 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	16429.7	1	16429.7	305
Residual	2744.8	51	53.8195	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	1.7847	2.096	0.851	0.3986
NumberOfCr...	3.70089	0.2118	17.5	≤ 0.0001

How Many Rooms Can x Crews Clean?

Regression Display

Dependent variable is: **RoomsClean**
No Selector
54 total cases of which 1 is missing
R squared = 85.7% R squared (adjusted) = 85.4%
s = 7.336 with 53 - 2 = 51 degrees of freedom

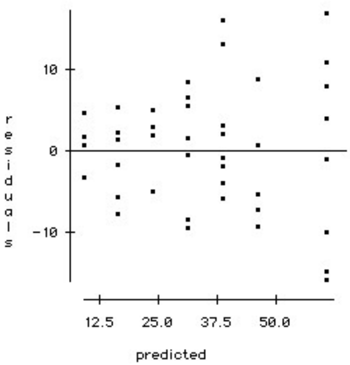
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	16429.7	1	16429.7	305
Residual	2744.8	51	53.8195	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1.7847	2.096	0.851	0.3986
NumberOfCr...	3.70089	0.2118	17.5	≤ 0.0001

$$\widehat{RoomsCleaned} = 3.70 \cdot NumCrews + 1.78$$

How Many Rooms Can x Crews Clean?

Residual Plot

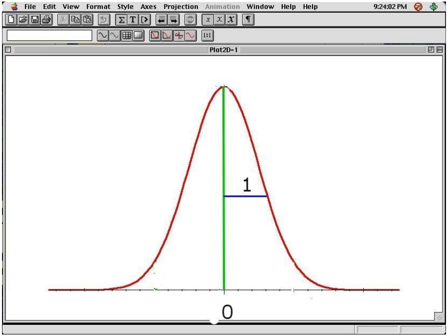


How Many Rooms Can x Crews Clean?

There are important deviations from the the assumptions of an ideal linear regression model here.

Normal Distribution Formula

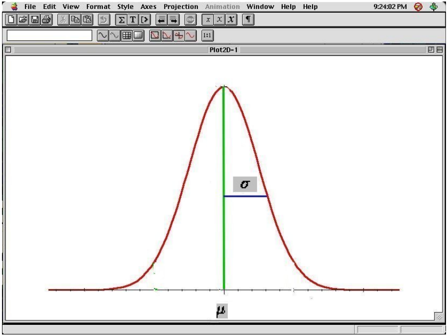
N(0,1)



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

Normal Distribution Formula

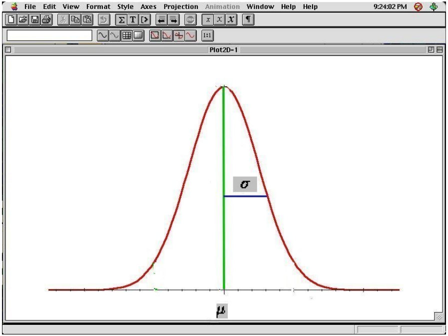
$N(\mu, \sigma)$



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

Normal Distribution Formula

$N(\mu, \sigma)$



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

These formulas and the following argument are far above the basic level of our course.

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$
$\mu = m$ and $\sigma = \sqrt{.5m}$.

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Why Normal Out of Binomial?

Math 1710
Class 22

V1

Last Time

Regression

R^2

Regression Conditions

Regression Cautions

1970 Draft Lottery

Gas Chromatography

Cleaning Crews Example

Proof of Normal Approximation

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

The z-score of $m + k$ is $\frac{k\sqrt{2}}{\sqrt{m}}$.

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

The z-score of $m + k$ is $\frac{k\sqrt{2}}{\sqrt{m}}$.

So we want to show $a_k \sim ce^{-\frac{k^2}{m}}$.

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

So we want to show $a_k \sim ce^{-\frac{k^2}{m}}$.

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.
Strategy: Compare a_k to a_0 using the approximation $\ln(1 + x) \sim x$ for x small.

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

$$a_k = \frac{(2m)!(.5)^{2m}}{(m+k)!(m-k)!} = a_0 \frac{(m)(m-1)\dots(m-k+1)}{(m+k)(m+k-1)\dots(m+1)}$$

$$= a_0 \frac{(1)(1-\frac{1}{m})\dots(1-\frac{k-1}{m})}{(1+\frac{k}{m})(1+\frac{k-1}{m})\dots(1+\frac{1}{m})}$$

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

$$a_k = \frac{(2m)!(.5)^{2m}}{(m+k)!(m-k)!} = a_0 \frac{(m)(m-1)\dots(m-k+1)}{(m+k)(m+k-1)\dots(m+1)}$$

$$= a_0 \frac{(1)(1-\frac{1}{m})\dots(1-\frac{k-1}{m})}{(1+\frac{k}{m})(1+\frac{k-1}{m})\dots(1+\frac{1}{m})}$$

So using $\ln(1+x) \sim x$,

$$\ln a_k \sim \ln a_0 - 2\left(\frac{1}{m} + \frac{2}{m} + \dots + \frac{k-1}{m}\right) - \frac{k}{m}$$

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.
So using $\ln(1 + x) \sim x$,

$$\ln a_k \sim \ln a_0 - 2\left(\frac{1}{m} + \frac{2}{m} + \ldots + \frac{k-1}{m}\right) - \frac{k}{m}$$

But $1 + 2 + \ldots k - 1 = \frac{k(k-1)}{2}$, so

Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

So using $\ln(1 + x) \sim x$,

$$\ln a_k \sim \ln a_0 - 2\left(\frac{1}{m} + \frac{2}{m} + \ldots + \frac{k-1}{m}\right) - \frac{k}{m}$$

But $1 + 2 + \ldots k - 1 = \frac{k(k-1)}{2}$, so

$$\ln a_k \sim \ln a_0 - \frac{k^2}{m}$$

as desired.