Regression

Cautions

1970 Draft Lottery

Gas Chromatography

Cleaning Crews

Example

Proof of Normal Approximation

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

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

\(-1 \leq r \leq 1\) (\(\pm 1\) 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

\[ r \text{ is strongly affected by outliers. Consider 9 points } (x=0, 1, \ldots, 8 \text{ on the line } y = 2x + 3 \text{ plus one outlier when } x = 9. \]

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>9</td>
<td>21 + \epsilon_9</td>
</tr>
</tbody>
</table>
$r$ is strongly affected by outliers.
Correlation Properties

$r$ is strongly affected by outliers.

**Correlation vs. size of $\epsilon_9$.**

<table>
<thead>
<tr>
<th>$\epsilon_9$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>-10</td>
<td>0.853</td>
</tr>
<tr>
<td>10</td>
<td>0.944</td>
</tr>
<tr>
<td>-5</td>
<td>0.968</td>
</tr>
<tr>
<td>-10</td>
<td>0.853</td>
</tr>
<tr>
<td>-15</td>
<td>0.663</td>
</tr>
<tr>
<td>-20</td>
<td>0.455</td>
</tr>
<tr>
<td>-25</td>
<td>0.275</td>
</tr>
<tr>
<td>5</td>
<td>0.981</td>
</tr>
<tr>
<td>15</td>
<td>0.904</td>
</tr>
<tr>
<td>20</td>
<td>0.866</td>
</tr>
<tr>
<td>25</td>
<td>0.834</td>
</tr>
</tbody>
</table>
Correlation Properties

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

samples from independent RV’s ⇒ \( r \sim 0 \)
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. \( \rho = .99 \)

\[ \Rightarrow \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 \(y = c_0 + c_1 x\)
Given any point \((x_i, y_i)\)
and any line \(y = c_0 + c_1x\)
we can use the line to get a predicted value

\[ \hat{y}_i = c_0 + c_1x_i \]
What does Best Fitting Mean?

Given any point \((x_i, y_i)\) and any line \(y = c_0 + c_1x\) we can use the line to get a predicted value

\[
\hat{y}_i = c_0 + c_1x_i
\]

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 $\sum d_i^2$ of the squares of the vertical distances.
What does Best Fitting Mean?

Best fitting means the line minimizing the sum $\Sigma d_i^2$ of the squares of the vertical distances.

Another try: Minimize $\Sigma d_i$.

But cancellation would be a problem.
What does Best Fitting Mean?

Best fitting means the line minimizing the sum $\Sigma d_i^2$ 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 $\sum d_i^2$ of the squares of the vertical distances.

Answer: We’ll show the best fitting line $\hat{y} = b_1x + b_0$ 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 $\Sigma d_i^2$ 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 \( r^2 \).”
“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:

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

\[
r^2 = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}
\]

More meaningful is the companion statement about residuals:

\[
\text{Var}(d_i) = (1 - r^2) \cdot \text{Var}(y_i)
\]
"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

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

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

\[
= \Sigma((y_i - \bar{y}) - c(x_i - \bar{x}) - d)^2
\]

\[
= \Sigma(y_i - \bar{y})^2 - 2c(x_i - \bar{x})(y_i - \bar{y}) + c^2(x_i - \bar{x})^2
\]

\[
+ \Sigma d^2 + 2cd(x_i - \bar{x}) - 2d(y_i - \bar{y})
\]
Proof of Formula for Line of Regression

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

\[\sum(y_i - \bar{y})^2 - 2c(x_i - \bar{x})(y_i - \bar{y}) + c^2(x_i - \bar{x})^2 + \sum d^2 + 2cd(x_i - \bar{x}) - 2d(y_i - \bar{y})\]

Use the definitions of \(s_x^2\), \(s_y^2\), and \(r\). Also use the fact that \(\sum(x_i - \bar{x}) = 0\) and \(\sum(y_i - \bar{y}) = 0\). 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 \((x_i, y_i), \; i = 1 \ldots n\)
- No Outliers
- Straight Enough Condition
Regression Conditions

Textbook:
- (Paired) Quantitative Variables $(x_i, y_i)$, $i = 1 \ldots n$
- 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 \( \epsilon_i \) independent of each other and all following \( N(0, \sigma) \) for some constant standard deviation \( \sigma \).
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 \( \epsilon_i \) independent of each other and all following \( N(0, \sigma) \) for some constant standard deviation \( \sigma \).

There’s no requirement that the \( x \) values be random.
Regression Cautions

Extrapolation is Dangerous

![Extrapolation Chart]

MY HOBBY: EXTRAPOLATING

As you can see, by late next month you'll have over four dozen husbands. Better get a bulk rate on wedding cake.
Regression Cautions

Extrapolation is Dangerous

As you can see, by late next month you'll have over four dozen husbands. Better get a bulk rate on wedding cake.

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

<table>
<thead>
<tr>
<th>No Selector</th>
<th>Draft_No.</th>
<th>Day_of_year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.000</td>
<td>-0.226</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

1970 Draft Lottery

Gas Chromatography

Cleaning Crews Example

Proof of Normal Approximation
Was it Fair?

Correlation Display

**Pearson Product-Moment Correlation**

No Selector

<table>
<thead>
<tr>
<th>Draft_No.</th>
<th>Draft_</th>
<th>Day_of_</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draft_No.</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Day_of_year</td>
<td>-0.226</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2.75907e6</td>
<td>1</td>
<td>2.75907e6</td>
<td>3.39e4</td>
</tr>
<tr>
<td>Residual</td>
<td>1465.53</td>
<td>18</td>
<td>81.4184</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-14.4107</td>
<td>2.614</td>
<td>-5.51</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>amount</td>
<td>46.6287</td>
<td>0.2533</td>
<td>184</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>
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.)
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
How Many Rooms Can x Crews Clean?

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

Percentile 25

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Count</strong></td>
<td>53</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>8.67925</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>StdDev</strong></td>
<td>4.80294</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>14</td>
</tr>
<tr>
<td><strong>Interquartile Range</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>Lower ith %tile</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>Upper ith %tile</strong></td>
<td>12</td>
</tr>
</tbody>
</table>
How Many Rooms Can x Crews Clean?

RoomsCleaned Summary

Summary of
No Selector
54 total cases of which 1 is missing
Percentile 25

Count 53
Mean 33.9057
Median 35
StdDev 19.2026
Range 72
InterQuartile Range 27.5
Lower 1st Quartile 18.75
Upper 1st Quartile 46.25
How Many Rooms Can x Crews Clean?

Scatterplot with Regression Line

[Diagram showing a scatterplot with a regression line, indicating the relationship between the number of crews and the number of rooms cleaned.]
# 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

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>16429.7</td>
<td>1</td>
<td>16429.7</td>
<td>305</td>
</tr>
<tr>
<td>Residual</td>
<td>2744.8</td>
<td>51</td>
<td>53.8195</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.7847</td>
<td>2.096</td>
<td>0.851</td>
<td>0.3986</td>
</tr>
<tr>
<td>Number Of Cr...</td>
<td>3.70089</td>
<td>0.2118</td>
<td>17.5</td>
<td>$\leq 0.0001$</td>
</tr>
</tbody>
</table>
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%  \quad R squared (adjusted) = 85.4%

s = 7.336 \quad with \quad 53 - 2 = 51 \quad degrees \quad of \quad freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>16429.7</td>
<td>1</td>
<td>16429.7</td>
<td>305</td>
</tr>
<tr>
<td>Residual</td>
<td>2244.8</td>
<td>51</td>
<td>53.8195</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.7847</td>
<td>0.22926</td>
<td>0.851</td>
<td>0.3986</td>
</tr>
<tr>
<td>Number Of Cr...</td>
<td>3.70089</td>
<td>0.2118</td>
<td>17.5</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>
How Many Rooms Can x Crews Clean?

Regression Display

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>16429.7</td>
<td>1</td>
<td>16429.7</td>
<td>305</td>
</tr>
<tr>
<td>Residual</td>
<td>2244.8</td>
<td>51</td>
<td>53.8195</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.7847</td>
<td>0.851</td>
<td>0.3986</td>
</tr>
<tr>
<td>Number Of Cr...</td>
<td>3.70089</td>
<td>17.5</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

$$\text{RoomsCleaned} = 3.70 \cdot \text{NumCrews} + 1.78$$
How Many Rooms Can x Crews Clean?

Residual Plot

Proof of Normal Approximation
How Many Rooms Can x Crews Clean?

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

\[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \]
Normal Distribution Formula

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Normal Distribution Formula

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

These formulas and the following argument are far above the basic level of our course.
Suppose $X \sim Binom(2m, .5)$
Why Normal Out of Binomial?

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

$\mu = m$ and $\sigma = \sqrt{.5m}$. 
Why Normal Out of Binomial?

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

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Why Normal Out of Binomial?

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

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

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$
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 $\frac{k\sqrt{2}}{\sqrt{m}}$. 
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 $a_k \sim ce^{-\frac{k^2}{m}}$. 
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. $\ln a_k \sim \ln c - \frac{k^2}{m}$.
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. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

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

$$= a_0 \frac{(1)(1-\frac{1}{m})\ldots(1-\frac{k-1}{m})}{(1+k/m)(1+\frac{k-1}{m})\ldots(1+\frac{1}{m})}$$
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. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

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

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

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

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