

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 10

What is the Distribution of the Residuals?

Chris A. Mack
Adjunct Associate Professor

<http://www.lithoguru.com/scientist/statistics/>

© Chris Mack, 2016 Data to Decisions 1

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Assumptions in OLS Regression

1. ε is a random variable that does not depend on x (i.e., the model is perfect, it properly accounts for the role of x in predicting y)
2. $E[\varepsilon_i] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All ε_i are independent of each other (uncorrelated for the population, but not for a sample)
4. All ε_i have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\varepsilon \sim N(0, \sigma^2)$ (the residuals, and thus the y_i , are normally distributed)
6. The values of each x_i are known exactly

© Chris Mack, 2016 Data to Decisions 2

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Central Limit Theorem

- If an error is the sum of many small, independent error sources, the total error will be about Normally distributed
 - Since this situation is not uncommon, errors frequently are about Normal
- The central limit theorem is *not* a law of physics – there is no guarantee it will apply
 - It is often hard to predict what the error distribution will be

© Chris Mack, 2016 Data to Decisions 3

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

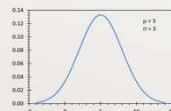
Normal Distribution

- Also called the Gaussian distribution

pdf $\rightarrow f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = N(\mu, \sigma)$

cdf $\rightarrow F_X(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x-\mu}{\sqrt{2}\sigma} \right) \right]$

$E[X] = \mu$
 $\operatorname{var}[X] = \sigma^2$




© Chris Mack, 2016 Data to Decisions 4

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD


Departures from Normality

Skew



Skewed right Nor Skewed left

Kurtosis
(how heavy are the tails?)



Heavy tails: extreme values are more likely
Light tails: extreme values are less likely

© Chris Mack, 2016 Data to Decisions 5

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Departures from Normality

- Small departures from normality have very little impact on regression results
- Large departures from normality can bias the regression results and decrease efficiency (make the true confidence intervals about the parameters and predictions larger)
 - Confidence intervals calculated under the assumption of normality will be wrong

© Chris Mack, 2016 Data to Decisions 6

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

How to Determine the pdf?

- **Density estimation:** the construction of an estimate, based on observed data, of an unobservable underlying probability density function (pdf).
- There are several ways to estimate the pdf of a variable
 - Histograms, kernel density estimation (smoothing)
 - Q-Q plots (normal probability plots)
 - Moment tests
 - Shapiro-Wilk test for normality
 - Many, many others

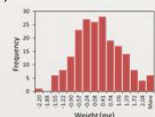
© Chris Mack, 2016 Data to Decisions 7

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Histograms

- **Histogram:** a graphical representation of the distribution of numerical data, plotting frequency (count) versus bins
 - A qualitative tool, hard to interpret
 - No objective way to find best bin size and bin start value
 - # bins = \sqrt{n} , # bins = $\log_2 n + 1$, # bins = $2n^{1/3}$
 - Relative uncertainty of each bin frequency is proportional to $1/\sqrt{\text{count}}$



© Chris Mack, 2016 Data to Decisions 8

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Lecture 10: What have we learned?

- Does the central limit theorem guarantee that error distributions are Normal?
- What are two common ways an actual distribution departs from Normality?
- What are the problems with using a histogram to assess distributions?

© Chris Mack, 2016 Data to Decisions 9