Basic Ideas in Probability and Statistics for Experimenters: Part II: Quantitative Work, Regression

He uses statistics as a drunken man uses lamp-posts – for support rather than for illumination ... A. Lang

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute
shivkuma@ecse.rpi.edu
http://www.ecse.rpi.edu/Homepages/shivkuma

Based in part upon slides of Prof. Raj Jain (OSU), Bill Foley (RPI)
Overview

- Quantitative examples of sample statistics, confidence interval for mean
- Introduction to Regression

- Also do the informal quiz handed out
- Reference: Chap 12, 13 (Jain), Chap 2-3 (Box, Hunter, Hunter),
- [http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html](http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html)
- Regression Applet: [http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html](http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html)
- Tool: R-project (free statistical package)
  - [http://www.r-project.org/](http://www.r-project.org/)
Independence

- $P(x \cap y) = \frac{1}{18}(2x + y)$ for $x = 1, 2$; and $y = 1, 2$
- and zero otherwise. Are the variables $X$ and $Y$ independent? Can you speculate why they are independent or dependent?

[Hint: $P(X)$ and $P(Y)$ can be formed by summing the above distribution (aka a joint distribution), since the combinations for specific values of $x$ and $y$ are mutually exclusive]
Independence

- \( P(x \cap y) = \frac{1}{30}(x^2 \cdot y) \) for \( x = 1,2; \) and \( y = 1,2,3 \)
  and zero otherwise. Are the variables \( X \) and \( Y \) independent? Can you speculate why they are independent or dependent?
Recall: Sampling Distribution

Uniform distribution looks **nothing** like bell shaped (gaussian)!
Large spread ($\sigma$)!

But the sampling distribution looks gaussian with smaller Standard deviation!

Sample mean $\sim N(\mu, \sigma/(n)^{0.5})$

i.e. the standard deviation of the sample mean (aka **standard error**) decreases with larger samples ($n$)

Shivkumar Kalyanaraman
Confidence Interval

- Sample mean: $\bar{x} \sim N(\mu, \sigma/(n)^{0.5})$
- The $100(1-\alpha)\%$ confidence interval is given by (if $n > 30$):
  \[
  \{\bar{x} - z_{(1-\alpha/2)} \frac{s}{(n)^{0.5}}, \bar{x} + z_{(1-\alpha/2)} \frac{s}{(n)^{0.5}}\} 
  \]
- $z_\beta \sim N(0, 1)$; i.e. it is the unit normal distribution
- $P\{(y - \mu)/\sigma \leq z_\beta\} = \beta$
- Eg 90% CI: $\{\bar{x} - z_{(0.95)} \frac{s}{(n)^{0.5}}, \bar{x} + z_{(0.95)} \frac{s}{(n)^{0.5}}\}$
- Refer to z-tables on pg 629, 630 (table A.2 or A.3)

\[ \text{FIGURE 2.12. Tail areas of the normal distribution.} \]
Meaning of Confidence Interval

http://www.math.csusb.edu/faculty/stanton/m262/confidence_means/confidence_means.html

FIGURE 13.1 Meaning of a confidence interval.
Ex: Sample Statistics, Confidence Interval

- Given: n=32 random RTT samples (in ms):
  {31, 42, 28, 51, 28, 44, 56, 39, 39, 27, 41, 36, 31, 45, 38, 29, 34, 33,
   28, 45, 49, 53, 19, 37, 32, 41, 51, 32, 39, 48, 59, 42}

1. Find: sample mean (xbar), median, mode, sample standard deviation (s), C.o.V., SIQR and 90% confidence interval (CI) & 95% CI for the population mean

2. Interpret your statistics qualitatively. I.e. what do they mean?

Hint: Refer to the formulas in pg 197 and pg 219 of Jain’s text (esp for s). Latter is reproduced in one of the slides
Box 13.1 Confidence Intervals

1. Given a sample \( x_1, x_2, \ldots, x_n \) of \( n \) observations:
\[ \bar{x} = \text{sample mean}; s = \text{sample standard deviation} \]

(a) Standard error of the sample mean: \( \sigma_x = \frac{s}{\sqrt{n}} \)
(b) 100(1 - \alpha)\% two-sided confidence interval for the mean:
\[ \bar{x} \pm z_{1-\alpha/2}s/\sqrt{n} \]
If \( n \leq 30 \):
\[ \bar{x} \pm t_{n-1-\alpha/2} \frac{s}{\sqrt{n}} \]
(c) 100(1 - \alpha)\% one-sided confidence interval for the mean:
\[ (\bar{x} - z_{1-\alpha}s/\sqrt{n}, \infty) \text{ or } (\frac{\bar{x} - t_{n-1-\alpha/2}s}{\sqrt{n}}, \infty) \]
If \( n \leq 30 \):
\[ (\bar{x} - s\sqrt{\frac{2}{n}}, \infty) \text{ or } (\bar{x} - t_{n-1-\alpha/2}s/\sqrt{n}, \bar{x}) \]

2. To compare two systems using unpaired observations:

(a) The standard error of the mean difference: \( \sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} \)
(b) The effective number of degrees of freedom:
\[ \nu = \frac{(s_a^2/n_a + s_b^2/n_b)^2}{\frac{1}{n_a + 1}\left(\frac{s_a^2}{n_a}\right)^2 + \frac{1}{n_b + 1}\left(\frac{s_b^2}{n_b}\right)^2 - 2} \]
(c) The confidence interval for the mean difference:
\[ (\bar{x}_a - \bar{x}_b) \pm t_{\nu, 1-\alpha/2} \sigma_{\bar{x}-\bar{x}} \]

3. If \( n_1 \) of the \( n \) observations belong to a certain class, the following statistics can be reported for the class:

(a) Proportion of the observations in the class: \( p = \frac{n_1}{n} \)
(b) 100(1 - \alpha)\% two-sided confidence interval for the proportion\(^1\):
\[ p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \]
(c) 100(1 - \alpha)\% one-sided confidence interval for the proportion\(^1\):
\[ \left( p, p + z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}} \right) \text{ or } \left( p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}, p \right) \]

\(^1\) Only for samples from normal populations.
\(^2\) Provided \( np \geq 10 \).
**t-distribution:** for confidence intervals given few samples (6 <= n < 30)

- Idea: t-distribution with n-1 degrees of freedom approximates normal distribution for larger n (n >= 6).

- t-distribution is a poor approximation for lower degrees of freedom (i.e. smaller number of samples than 6!).

![Figure 2.13. The t distribution for v = 1, 9, and \( \infty \).](image)
t-distribution: confidence intervals

- The 100(1-\(\alpha\))% confidence interval is given by (if \(n \leq 30\)):

\[
\{\bar{x} - t_{\{1-\alpha/2, n-1\}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\{1-\alpha/2, n-1\}} \frac{s}{\sqrt{n}}\}
\]

- Use t-distribution tables in pg 631, table A.4

- Eg: for \(n = 7\), 90% CI:

\[
\{\bar{x} - t_{\{0.95, 6\}} \frac{s}{\sqrt{7}}, \bar{x} + t_{\{0.95, 6\}} \frac{s}{\sqrt{7}}\}
\]

*FIGURE 13.2* The ratio \((\bar{x} - \mu) / (s/\sqrt{n})\) for samples from normal populations follows a \(t(n - 1)\) distribution.
Confidence Interval with few samples

- Given: n=10 random RTT samples (in ms):
  \{31, 42, 28, 51, 28, 44, 56, 39, 39, 27\}

- 1. Find: sample mean (xbar), sample standard deviation (s) and 90% confidence interval (CI) & 95% CI for the population mean

- 2. Interpret this result relative to the earlier result using the normal distribution.

- Hint: Refer to the formulas in pg 197 and pg 219 of Jain’s text (esp for s).
Linear Regression

- Goal: determine the relationship between two random variables $X$ and $Y$.
- Example: $X =$ height and $Y =$ weight of a sample of adults.

- Linear regression attempts to explain this relationship with a straight line fit to the data, i.e. a linear model.

- The linear regression model postulates that
  \[ Y = a + bX + e \]

- Where the "residual" or "error" $e$ is a random variable with mean $= 0$.

- The coefficients $a$ and $b$ are determined by the condition that the sum of the square residuals (i.e. the "energy" of residuals) is as small as possible (i.e. minimized).
Linear Equations

\[ Y = mX + b \]

- \( b \) = Y-intercept
- \( m \) = Slope

\[ \text{Change in Y} \quad \text{Change in X} \]

\[ b = Y\text{-intercept} \]

High School Teacher

Shivkumar Kalyanaraman

Rensselaer Polytechnic Institute

© 1984-1994 T/Maker Co.
1. Relationship Between Variables Is a Linear Function

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]
Population & Sample Regression Models

Population

Unknown Relationship

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

Random Sample

\[ Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i \]
Population Linear Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

\[ E(Y) = \beta_0 + \beta_1 X_i \]

where \( \varepsilon_i \) is the random error.
Sample Linear Regression Model

\[ Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i \]

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

\( \hat{\varepsilon}_i \) = Random error

Observed value

Unsampled observation
Demo: Regression Applet

- [http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html](http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html)

**FIGURE 14.1** Good and bad regression models.
Regression Theory

- **Model**: \( y_{est} = b_0 + b_1x_i \)
- **Error**: \( e_i = y_i - y_{est} \)
- **Sum of Squared Errors (SSE)**: \( \Sigma e_i^2 = \Sigma(y_i - b_0 + b_1x_i)^2 \)
- **Mean error** (\( e_{avg} \)): \( \Sigma e_i = \Sigma(y_i - b_0 + b_1x_i) \)

- **Linear Regression problem**: Minimize: \( \text{SSE} \) Subject to the constraint: \( e_{avg} = 0 \)

- **Solution**: (i.e. regression coefficients)
  - \( b_1 = \frac{s_{xy}^2}{s_x^2} = \frac{\Sigma xy - x_{avg} y_{avg}}{\Sigma x^2 - n(x_{avg})^2} \)
  - \( b_0 = y_{avg} - b_1 x_{avg} \)
Least Squares

1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum
   - But Positive Differences Off-Set Negative

   \[ \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2 \]

2. LS Minimizes the Sum of the Squared Differences (SSE)
Random Error Variation

1. Variation of Actual $Y$ from Predicted $Y$
2. Measured by Standard Error of Regression Model
   - Sample Standard Deviation of $\varepsilon$, $s^*$
3. Affects Several Factors
   - Parameter Significance
   - Prediction Accuracy
Measures of Variation in Regression

1. Total Sum of Squares (SS$_{yy}$)
   - Measures Variation of Observed $Y_i$ Around the Mean $\bar{Y}$

2. Explained Variation (SSR)
   - Variation Due to Relationship Between $X$ & $Y$

3. Unexplained Variation (SSE)
   - Variation Due to Other Factors
Variation Measures

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

Total sum of squares

$$(Y_i - \bar{Y})^2$$

Unexplained sum of squares

$$(Y_i - \hat{Y}_i)^2$$

Explained sum of squares

$$(\hat{Y}_i - \bar{Y})^2$$
Coefficient of Determination (R-squared)

1. Proportion of Variation ‘Explained’ by Relationship Between $X$ & $Y$

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \sum_{i=1}^{n} (Y_i - \hat{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

$0 \leq r^2 \leq 1$
Coefficient of Determination

Examples

\[ r^2 = 1 \]

\[ r^2 = 0.8 \]

\[ r^2 = 1 \]

\[ r^2 = 0 \]
Linear Regression Assumptions

- 1. Mean of Probability Distribution of Error Is 0
- 2. Probability Distribution of Error Has Constant Variance
- 3. Probability Distribution of Error is Normal
- 4. Errors Are Independent
Practical issues: Check linearity hypothesis with scatter diagram!

**FIGURE 14.6** Possible patterns of scatter diagrams.

(a) Linear  (b) Multilinear

(c) Outlier  (d) Nonlinear
Practical issues: Check randomness & zero mean hypothesis for residuals!

FIGURE 14.7 Possible patterns of residual versus predicted response graphs.
Practical issues: Does the regression indeed explain the variation?

- Coefficient of determination ($R^2$) is a measure of the value of the regression (i.e. variation explained by the regression relative to simple second order statistics).

- But it can be misleading if scatter plot is not checked.

![Graphs showing examples of data with high coefficient of determination](image)

FIGURE 15.5 Examples of data that may give high coefficient of determination, but the linear model obtained may not represent the system correctly.

Shivkumar Kalyanaraman
Non-linear regression: Make linear through transformation of samples!

FIGURE 15.2 Standard deviation versus mean response graphs can be used to determine the transformation required.
CIs for Regression & Predictions

\[ Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \]