"Around" the Normal Distribution

Sums of normals:

If

\[ Y_i \sim N(\mu_i, \sigma_i^2), \hspace{1em} 1 \leq i \leq n, \]

then

\[ S = \sum_{i=1}^{n} Y_i \sim N\left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right). \]

Sum of normals is normal

Log-Normal random variables

If \( X \sim N(\mu, \sigma^2) \), then

\[ Y = e^X \sim \ln N(\mu, \sigma^2) \]

and conversely if \( Y \sim \ln N(\mu, \sigma^2) \), then

\[ X = \ln(Y) \sim N(\mu, \sigma^2) \]

Note \(-\infty < X < \infty\) and \(0 < Y < \infty\).

If

\[ Q = \prod_{i=1}^{m} Y_i, \quad Y_i \sim \ln N, \]

then \( Q \) is also \( \ln N \):

\[ \ln Q = \sum_{i=1}^{m} \ln Y_i \Rightarrow \ln Q \sim N \]

\[ Q = e^{\ln Q} \sim \ln N. \]
Artists rendition: not mean of Y

\[ f_Y(\ln N|\mu = 0, \sigma = 5) \]

Note if \( \mu \gg \sigma \) then \( \ln N \sim N \).

Pomelo estimation:

\[ Q = \rho \cdot \frac{1}{\sqrt{3}} \cdot a \cdot b \cdot c \]

Mass density principle axes of ellipsoid

If \( \rho, a, b, c \) \( \ln N \)

\[ \ln N \leftarrow \text{then} \]

To "test": compare histogram of \( n = 53 \)

\[ \ln Q - \text{test} \]

To histograms from standard normal \( Z \sim N(0,1) \).

Central Limit Theorem (one version)

Let \( X_i, 1 \leq i \leq n \), be iid r.v. - \( f_X \)

with mean \( \mu \) and variance \( \sigma^2 \)

and

\[ \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]

be the sample mean;

then as \( n \to \infty \)

\[ \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to Z \]

\[ \Phi(z) \]
Example: Bernoulli

Let $X_i, 1 \leq i \leq n$, be i.i.d. $f_{\text{Bernoulli}}$, mean $\Theta$ and variance $\Theta(1-\Theta)$,

and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ (fraction heads) be the sample mean,

then as $n \to \infty$

$$P\left(\frac{\bar{X}_n - \Theta}{\sqrt{\frac{\Theta(1-\Theta)}{n}}} \leq z\right) \to \Phi(z).$$

"2.086" accuracy criterion: $n \Theta > 5$ and $n(1-\Theta) > 5$

Estimation: Bernoulli

Estimator for $\Theta$:

Recall if $X_i, 1 \leq i \leq n$, are i.i.d. $f_{\text{Bernoulli}}$, and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sample mean,

then $E(\bar{X}_n) = \Theta$ and $E((\bar{X}_n - \Theta)^2) = \Theta(1-\Theta)/n$.

 estimator for $\Theta$  good estimator for $\Theta$
Thus define
\[ \hat{\Theta}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]
as our estimator for \( \Theta \), and
\[ \hat{\Theta}_n = \tilde{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]
as our estimate for \( \Theta \).

Note: \( \Theta \) parameter, \( \hat{\Theta}_n \) estimator, \( \tilde{\Theta}_n \) estimate.

Confidence Interval:
2-sided, normal approximation
\[
P\left( \frac{\hat{\Theta}_n - \Theta}{\sqrt{\frac{\Theta(1-\Theta)}{n}}} < z \right) = \Phi(z)
\]
Recall for large \( n \)
\[
P\left( -z < \frac{\hat{\Theta}_n - \Theta}{\sqrt{\frac{\Theta(1-\Theta)}{n}}} < z \right) = \Phi(z) - \Phi(-z)
\]
\[
= \Phi(z) - (1 - \Phi(z))
\]
\[
= 2\Phi(z) - 1.
\]

Choose
\[ 2\Phi(z_Y) - 1 = \gamma \] confidence level

Hence
\[ \Phi(z_Y) = \frac{1 + \gamma}{2} \]
\[ z_Y = \frac{z}{\sqrt{1 + \gamma}} \] quantile of \( \Phi \)

Or
\[ z_Y = \frac{z}{\sqrt{\gamma(1-\gamma)}} \]

Note \( \gamma = 0 \Rightarrow z_Y = 0 \) (median), and
\[ \gamma \to 1 \Rightarrow z_Y \to \infty, \] and
\[ \gamma = 0.95 \Rightarrow z_Y \approx 1.96. \]
Define
\[
[\text{CI}]_n = \left[ \hat{\theta}_n - z_Y \sqrt{\frac{\hat{\sigma}^2}{n}} , \hat{\theta}_n + z_Y \sqrt{\frac{\hat{\sigma}^2}{n}} \right]
\]
and then (since \( \theta \) unknown) for \( n \) large

\[
[\text{CI}]_n = \left[ \hat{\theta}_n - z_Y \sqrt{\frac{\hat{\sigma}^2}{n}}, \hat{\theta}_n + z_Y \sqrt{\frac{\hat{\sigma}^2}{n}} \right]
\]

Hence

\[
P\left( \theta \text{ is inside } [\text{CI}]_n \right) = \gamma
\]

Note: \( \theta \) is a deterministic parameter, whereas 

\([\text{CI}]_n \) is a random interval.

In practice: choose \( n, \gamma \rightarrow 2 \gamma \)

\[
\text{sample } x_1, x_2, \ldots, x_n;
\]

compute estimate for \( \theta \),

\[
\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i;
\]

compute \([\text{CI}]_n \),

\[
[\text{CI}]_n = \left[ \hat{\theta}_n - z_Y \sqrt{\frac{\hat{\sigma}^2}{n}}, \hat{\theta}_n + z_Y \sqrt{\frac{\hat{\sigma}^2}{n}} \right]
\]

HalfLength\_\( \theta_n \) = \( z_Y \sqrt{\frac{\hat{\sigma}^2}{n}} \),

RelErr\_\( \theta_n \) = \( \frac{\text{HalfLength\_}\theta_n}{\hat{\theta}_n} \) = \( z_Y \sqrt{\frac{1-\theta_n}{\hat{\theta}_n n}} \).

Note:

as \( \gamma \rightarrow 1, z_Y \rightarrow \infty \), and \( \text{HalfLength}_{\theta, n} \rightarrow \infty \)

more confidence \( \Rightarrow \) less accuracy

as \( n \rightarrow \infty \), \( \text{HalfLength}_{\theta, n} \rightarrow 0 \) but SLOWLY

more samples \( \Rightarrow \) more accuracy,

as \( \theta (\hat{\theta}_n) \rightarrow 0 \), \( \text{RelErr}_{\theta, n} \rightarrow \infty \) for fixed \( n \)

rare event \( \Rightarrow \) less accuracy.

(Also require \( n \hat{\theta} > 5 \) and \( n(1-\hat{\theta}) \) for normal approximation.)

Frequentist interpretation:

\[
\gamma = 0.8
\]
\[
\theta = 5
\]

\[
\gamma = 0.95
\]
\[
\theta = 5
\]

50 experiments each with \( n = 100 \) coin flips.

Figure 10a: An example of confidence intervals for estimating the mean of a Bernoulli random variable (\( \theta = 0.5 \)) using 100 samples.
Some Applications of Bernoulli Estimation

The Birthmonth "Distribution"

Let

\[ X = \begin{cases} 
0 & \text{birthmonth [Jan-June]} \quad \text{probability } 1-\theta \\
1 & \text{birthmonth [July-Dec]} \quad \text{probability } \theta 
\end{cases} \]

Choose

\[ n = 51 \quad (2006 \text{ class size}) \]
\[ \theta = 0.95 \quad (\text{confidence level}) \]

Collect data: \( n = 51 \)

- birthmonth_number(\(i\)), \(1 \leq i \leq 12\),
- is \# occurrences of birthmonth \(i\),
- note \( \text{sum}(\text{birthmonth_number}) = 51 \).

Compute estimate for \( \theta \):

\[ \hat{\theta}_{n=51} = \frac{\text{sum}(\text{birthmonth_number}[7:12])}{51} \]
\[ = \frac{26}{51} = .5098 \]
\[ (= \frac{\sum x_i}{n}) \quad \hat{\theta}_{n} = 27.05 \quad (> 5) \checkmark \]

\[ n(1-\hat{\theta}_{n}) = 25.98 \quad (> 5) \checkmark \]

Calculate confidence interval for \( \theta \):

\[ Z_{0.95} = 1.96 \]

\[ 1.96 \sqrt{\frac{\hat{\theta}_{n}(1-\hat{\theta}_{n})}{n}} = .1872 \]

\[ [\hat{\theta}_{n}]_{\theta} = [ .5098 - .1872, .5098 + .1872 ] \]

\[ = [.3226, .6470] \]

Conclusion: If we are amongst the lucky 9,509/10,000 parallel universes,

\[ .3226 \leq \theta \leq .6470 \]

(and no reason to reject hypothesis \( \theta = \frac{1}{2} \)).
Other applications (same game)

Quality control:
\[
X = \begin{cases} 
0 & \text{part not up to spec} \quad \text{probability } 1-\theta \\
1 & \text{part up to spec} \quad \text{probability } \theta 
\end{cases}
\]
\[n = \#(\text{parts}) \text{ from large population inspected}\]
\[\hat{\theta}_n = \text{fraction of parts up to spec}.\]

Failure:
\[
X = \begin{cases} 
0 & \text{shut } \sigma \leq \sigma_{\text{max}} \quad \text{probability } 1-\theta \\
1 & \text{shut } \sigma > \sigma_{\text{max}} \quad \text{probability } \theta (\ll 1)
\end{cases}
\]
\[n = \#(\text{shuts}) \text{ from large population inspected}\]
\[\hat{\theta}_n = \text{fraction of shuts}.\]

Preferences:
\[\text{two choices: } A, B\]
\[X = \begin{cases} 
0 & \text{prefer } A \quad \text{probability } 1-\theta \\
1 & \text{prefer } B \quad \text{probability } \theta 
\end{cases}\]
\[n: \text{number of voters in survey sample (focus group)}\]
\[\hat{\theta}_n = \text{fraction of voters who prefer } B\]

Note if popular (simple majority) election, \(y\)
\[A \text{ wins } \rightarrow \frac{[c]}{[c]}, \frac{[c]}{[c]}, \frac{[c]}{[c]}, \frac{[c]}{[c]}, B \text{ wins } \rightarrow \frac{[c]}{[c]}, \frac{[c]}{[c]}, \frac{[c]}{[c]}, \frac{[c]}{[c]}, \text{"too close to call"} \]
2.086 Numerical Computation for Mechanical Engineers
Spring 2013

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.