"around" the
Normal Distribution

Sums of normals:

If

$$Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad 1 \le i \le n,$$

then

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

sum of normals is normal

Log-Normal random variables

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Y = e^X \sim \ln\mathcal{N}(\mu, \sigma^2)$$

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

$$X = \ln(Y) \sim \mathcal{N}(\mu, \sigma^2).$$

Note $-\infty < X < \infty$ and $0 < Y < \infty$.

If

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

then Q is also $\ln\mathcal{N}$:

$$\ln Q = \sum_{i=1}^{m} \underbrace{\ln Y_i}_{\text{normal}} \Rightarrow \ln Q \sim \mathcal{N}$$

$$\Downarrow$$

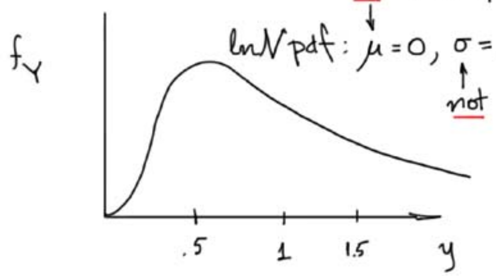
$$Q = e^{\ln Q} \sim \ln\mathcal{N}.$$

Artist's rendition:

lnN pdf: $\mu = 0$, $\sigma = .5$ $\quad e^X : X \sim \mathcal{N}(\mu, \sigma^2)$

not mean of Y

not std dev of Y



Note if $\mu \gg \sigma$ then $\ln\mathcal{N} \sim \mathcal{N}$.

Pomelo estimation:

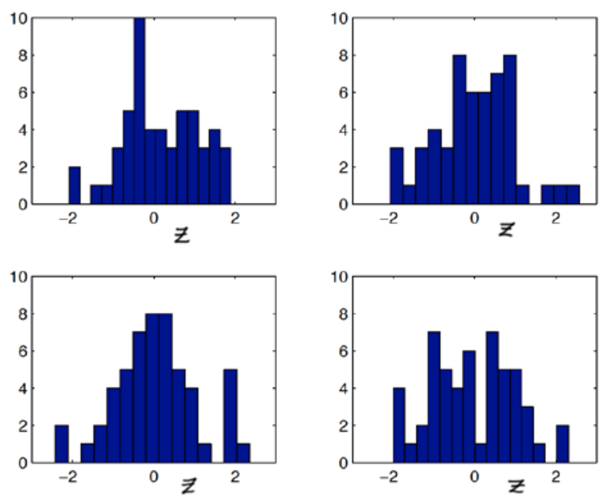
$$\underbrace{Q}_{\text{mass}} = \underbrace{\underbrace{\rho}_{\text{density}} \cdot \tfrac{4\pi}{3} \cdot \underbrace{a \cdot b \cdot c}_{\text{principle axes of ellipsoid}}}$$

$\ln\mathcal{N} \longleftarrow$ then, if ρ, a, b, c $\ln\mathcal{N}$

To "test": compare histogram of $\quad n = 53$

$$\frac{\ln Q - \mu_{est}}{\sigma_{est}}$$

to histograms from standard normal $Z \sim \mathcal{N}(0,1)$.



Central Limit Theorem (one version)

Let

X_i, $1 \le i \le n$, be i.i.d r.v. $\sim f_X$ with mean μ and variance σ^2,

and

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean;

then as $n \to \infty$

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z \right) \to \Phi(z) .$$

standard normal cdf

Example: Bernoulli binomial

Let

X_i, $1 \le i \le n$, be i.i.d. $\sim f_X^{\text{Bernoulli}}(x;\theta)$

$\Rightarrow$ mean θ and variance $\theta(1-\theta)$,

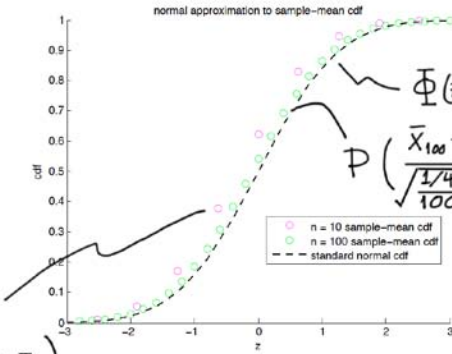
and

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

then as $n \to \infty$

$$P\left(\frac{\bar{X}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le z \right) \to \Phi(z). \qquad \approx$$

"2.086" accuracy criterion: $n\theta > 5$ AND $n(1-\theta) > 5$



$\Phi(z)$

$P\left(\frac{\bar{X}_{100} - \frac{1}{2}}{\sqrt{\frac{1/4}{100}}} \le z \right)$

$\theta = 0.5$

$P\left(\frac{\bar{X}_{10} - \frac{1}{2}}{\sqrt{\frac{1/4}{10}}} \le z \right)$

Estimation: Bernoulli

Estimator for θ:

Recall if

X_i, $1 \le i \le n$, are i.i.d. $f_X^{\text{Bernoulli}}(x;\theta)$

and

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean,

then

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

estimator for θ "good" estimator for θ

Thus define

$$\hat{\Theta}_n \equiv \bar{X}_n \left(= \frac{1}{n} \sum_{i=1}^{n} X_i \right)$$

as our <u>estimator</u> for θ, and

$$\hat{\theta}_n \equiv \bar{x}_n \left(= \frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

as our <u>estimate</u> for θ.

realization of

Note:

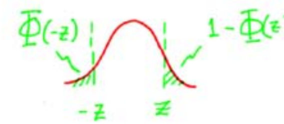
θ: parameter; $\hat{\Theta}_n$: estimator; $\hat{\theta}_n$: estimate.

deterministic — random variable — number

Confidence Interval: 2-sided, normal approximation

Recall for large n $\quad z > 0$

$$P\left(\frac{\hat{\Theta}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} < z \right) \approx \Phi(z)$$



so

$$P\left(-z \le \frac{\hat{\Theta}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le z \right) \approx \Phi(z) - \Phi(-z)$$

$$= \Phi(z) - (1 - \Phi(z))$$

$$= 2\Phi(z) - 1 .$$

Choose

$$2\Phi(z_\gamma) - 1 = \gamma \qquad \text{confidence level}$$

hence

$$\Phi(z_\gamma) = (1+\gamma)/2$$

or

$$z_\gamma = \tilde{z}_{(1+\gamma)/2} \quad \left(\tfrac{1+\gamma}{2} \text{ quantile of } \Phi \right)$$

Note $\gamma = 0 \Rightarrow z_\gamma = 0$ (median), and

$\gamma \to 1 \Rightarrow z_\gamma \to \infty$, and

$\gamma = 0.95 \Rightarrow z_\gamma \approx 1.96$.

Thus,

$$\underbrace{P\left(-z_\gamma \le \frac{\hat{\Theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le z_\gamma \right)} = \gamma$$

$$\Updownarrow$$

$$-z_\gamma \le \frac{\hat{\Theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \Rightarrow \theta \le \hat{\Theta} + z_\gamma \sqrt{\frac{\theta(1-\theta)}{n}}$$

$$\frac{\hat{\Theta}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le z_\gamma \Rightarrow \hat{\Theta}_n - z_\gamma \sqrt{\frac{\theta(1-\theta)}{n}} \le \theta$$

or

$$\underbrace{\hat{\Theta}_n - z_\gamma \sqrt{\frac{\theta(1-\theta)}{n}} \le \theta \le \hat{\Theta}_n + z_\gamma \sqrt{\frac{\theta(1-\theta)}{n}}}_{\text{with probability } \gamma} .$$

Define

$$[CI]_n^0 \equiv \left[\hat{\Theta}_n - z_\gamma \sqrt{\frac{\theta(1-\theta)}{n}},\ \hat{\Theta}_n + z_\gamma \sqrt{\frac{\theta(1-\theta)}{n}} \right]$$

and then (since θ unknown) $\qquad n$ large

$$[CI]_n \equiv \left[\hat{\Theta}_n - z_\gamma \sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n}},\ \hat{\Theta}_n + z_\gamma \sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n}} \right].$$

Hence

$$P(\theta \text{ is inside } [CI]_n) = \gamma \quad \approx$$

Note: θ is a deterministic parameter, whereas $[CI]_n$ is a random interval.

In practice: choose $n, \gamma \qquad \Rightarrow z_\gamma$

sample $x_1, x_2, \ldots, x_n$; $\qquad$ 0 or 1

compute estimate for θ,

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \, ;$$

compute $[ci]_n$,

$$[ci]_n = \left[\hat{\theta}_n - z_\gamma \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}},\ \hat{\theta}_n + z_\gamma \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \right],$$

$$\text{HalfLength}_{\theta;n} = z_\gamma \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}},$$

$$\text{RelErr}_{\theta;n} = z_\gamma \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \Big/ \hat{\theta}_n = z_\gamma \sqrt{\frac{1-\hat{\theta}_n}{\hat{\theta}_n \cdot n}}.$$

Note:

as $\gamma \to 1$, $z_\gamma \to \infty$, and $\text{HalfLength}_{\theta;n} \to \infty$

more confidence $\Rightarrow$ less accuracy;

as $n \to \infty$, $\text{HalfLength}_{\theta;n} \to 0$ $\quad$ but SLOWLY

more samples $\Rightarrow$ more accuracy,

as $\theta \ (\hat{\theta}_n) \to 0$, $\text{RelErr}_{\theta;n} \to \infty$ $\qquad$ fixed n

rare event $\Rightarrow$ less accuracy. $\dagger$

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

Frequentist interpretation:

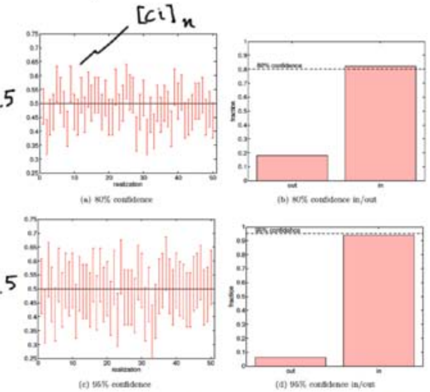
$[ci]_n$

$\gamma = 0.8$

$\theta = .5$

$\gamma = 0.95$

$\theta = .5$



50 experiments each with $n = 100$ coin flips

Figure 10.3: An example of confidence intervals for estimating the mean of a Bernoulli random variable ($\theta = 0.5$) using 100 samples.

$\theta = 0.5 \quad n = 100$

Some Applications
of
Bernoulli Estimation

The Birthmonth "Distribution"

Let

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

Choose

$n = 51$ (2.086 class size)

$\gamma = 0.95$ (confidence level).

Collect data: $n = 51$

birthmonth_number(i), $1 \le i \le 12$,
is #(occurrences of birthmonth i);

note sum(birthmonth_number) = 51.

Compute estimate for θ:

$\hat{\theta}_{n=51}$ = sum(birthmonth_number([7:12]))/51

$= 26/51 = .5098$

$\left(= \sum_{i=1}^{n} x_i / n \right)$.

$n\hat{\theta}_n = 27.05$ (>5) ✓

$n(1-\hat{\theta}_n) = 25.98$ (>5) ✓

Calculate confidence interval for θ: $z_{0.95} = 1.96$

$$1.96\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} = .1372$$

$$\Downarrow$$

$$[ci]_n = [.5098 - .1372, .5098 + .1372]$$

$$= [.3726, .6470].$$

Conclusion: if we are amongst the lucky 9,500/10,000 parallel universes,

$$\boxed{.3726 \le \theta \le .6470}$$

(and no reason to reject hypothesis $\theta = \frac{1}{2}$).

Other applications (same game)

Quality control:

$$X = \begin{cases} 0 & \text{part } \underline{\text{not up to spec}} & \text{probability } 1-\theta \\ 1 & \text{part } \underline{\text{up to spec}} & \text{probability } \theta \end{cases}$$

$n \equiv$ #(parts) from large population inspected

$\hat{\theta}_n \equiv$ fraction of parts up to spec.

Failure:

$$X = \begin{cases} 0 & \text{strut } \underline{\sigma \le \sigma_{max}} & \text{probability } 1-\theta \\ 1 & \text{strut } \underline{\sigma > \sigma_{max}} & \text{probability } \theta \ (\ll 1) \end{cases}$$

$n \equiv$ #(struts) from large population inspected

$\hat{\theta}_n$: #(cars in fleet) $\equiv$ #(repairs).

Preferences:

two choices: A, B

└ competing candidates, or products, or ...

$$X = \begin{cases} 0 & \underline{\text{prefer A}} & \text{probability } 1-\theta \\ 1 & \underline{\text{prefer B}} & \text{probability } \theta \end{cases}$$

n: number of voters in survey sample (focus group)

$\hat{\theta}_n$: fraction of voters who prefer B

Note if popular (simple majority) election, γ

A wins $[\mathrm{CI}]_n$ $[\mathrm{C}|\mathrm{I}]_n$ $[\mathrm{CI}]_n$ B wins.

.5 "too close to call"