An estimator $\tilde{\theta}$ is a function of random variables, i.e. g(Y1, Y2, …, Y_n)

If we know the distribution of each random variable Y_i, we can get a sampling distribution for $\tilde{\theta}$

- **Example**: $\tilde{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is an estimator
  - It is also a random variable
  - Has a corresponding distribution
- **Example**: $\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is an estimate
  - It is a function of the data at hand
  - Takes on a numerical value

- If $Y \sim G(\mu, \sigma)$ then $\bar{Y} \sim G(\mu, \sigma/\sqrt{n})$
- **Question**: If $\sigma = 2$ and $n = 50$, how often is the estimator $\bar{\mu} = \bar{Y}$ within 0.01 kg of the true mean $\mu$? What if $n = 100$?
- **Key idea**: understanding the sampling distribution of $\tilde{\theta}$ allows us to compute $P(|\tilde{\theta} - \theta| \leq d)$ for a given $d$

Ex. Suppose we want an estimator for the weight of geese.

The weight of goose i is stored in the dataset {Y1=y1, Y2=y2… }

How often is the estimator $\tilde{\mu} = \bar{Y}$ within 0.01 kg of the true mean μ?

↳ unknown

Essentially, we want

$$P(|\tilde{\mu} - \mu| \leq 0.01) = P(|\bar{Y} - \mu| \leq 0.01)$$

$$= P(-0.01 + \mu \leq \bar{Y} \leq 0.01 + \mu)$$

$$= P(-0.01 \leq \bar{Y} - \mu \leq 0.01)$$

$$= P\left(\frac{-0.01}{\sigma/\sqrt{n}} \leq z \leq \frac{0.01}{\sigma/\sqrt{n}}\right)$$

Plugging in values for σ and n allows us to calculate the probability.

As n gets larger, the denominator gets smaller, and each bound gets larger.

This means that there is a greater range of values for z to take on.

As such, the *probability of the estimator being within 0.01 of the true mean increases*.

Essentially, we want $P(|\tilde{\theta} - \theta| \leq d)$, for some choice of d, to be reasonably close to 1.

- *Remember:* both $\tilde{\theta}$ and θ are functions of the random variables
- Common choices of d are 0.05 or 0.01

# Confidence Intervals and Pivotal Quantities

Ex. Suppose we have a series of observations y1, y2, … y_n from a *Gaussian* sample; corresponding to the weights of geese around Waterloo. We want to estimate the true mean μ of geese in Waterloo.

- The MLE is $\bar{y}$ (sample mean) exact value $\hat{\theta}$
- The maximum likelihood *estimator* is $\bar{Y}$ = sum(y1, y_n) / n $\tilde{\theta}$
  must be a function

Then, an *interval estimate of* μ would be [$\bar{y}$ - 1.96σ/sqrt(n), $\bar{y}$ + 1.96σ/sqrt(n)]

- 1.96 is chosen because it leads to a nice number at the end

Suppose σ=2. Then, since Y is Gaussian, P(true mean μ is within the interval estimate) is equal to

$$P(\mu \in [\bar{Y} - 1.96 \cdot \frac{2}{\sqrt{n}}, \bar{Y} + 1.96 \cdot \frac{2}{\sqrt{n}}]$$

$$= P(\bar{Y} - 1.96 \cdot \frac{2}{\sqrt{n}} \leq \mu \leq \bar{Y} + 1.96 \cdot \frac{2}{\sqrt{n}})$$

$$= P(-1.96 \leq \underbrace{\frac{\bar{Y} - \mu}{2/\sqrt{n}}}_{\text{somehow???}} \leq 1.96) = P(-1.96 \leq Z \leq 1.96) = 0.95$$

In this case, $P(|\tilde{\theta} - \theta| \leq d) = 0.95$.

We cannot conclude that P(*interval estimate* of μ is within the interval) = 0.95 because the interval estimate of μ is a *constant*, not a random variable. It's either within the interval or not.

$P(\theta \in [L(y), U(y)]) = 0.95$ is FALSE

But we can conclude that P($\bar{Y}$ is within the interval) = 0.95

- For example, if we collect 10000 samples and use the same *interval estimator* in calculating an estimate for each, in 9500 of these samples, the interval estimate will be within the interval

As such, the above interval is a *95% confidence interval*. The *coverage probability* of the interval is 0.95

<u>Pivotal Quantities</u>

A pivotal quantity $Q = Q(Y_i, \theta)$ is a function of data Y and parameter θ such that Q is a random variable with known distribution

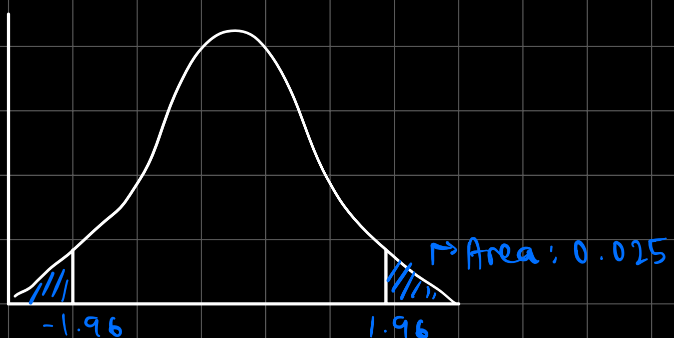
## Constructing Two-Sided Confidence Intervals from Pivotal Quantities

- Suppose we can rearrange the inequality
$$a \le Q(Y;\theta) \le b$$
as
$$L(Y) \le \theta \le U(Y)$$
- Then,
$$\begin{aligned} p &= P(a \le Q(Y;\theta) \le b) \\ &= P(L(Y) \le \theta \le U(Y)) \\ &= P(\theta \in [L(Y), U(Y)]) \end{aligned}$$

If $p = 0.95$, then $[L(Y), U(Y)]$ is a 95% confidence interval for the parameter $\theta$

And $P(L(Y) \le \theta \le U(Y)) = p$ is the *coverage probability* of this interval

Ex. Suppose we want a 95% confidence interval over $[\bar{y} - 1.96\sigma/\text{sqrt}(n)), \bar{y} + 1.96\sigma/\text{sqrt}(n))]$




Ex. Which is wider: 99% CI or 95% CI?

- 99% CI has cutoffs closer to the edges, since the area of the tails is smaller (0.005 vs 0.025)
- So it is wider, since the range of values $\theta$ can take on is wider

# Asymptotic Pivotal Quantities

Consider a random variable Y with a distribution other than G(0,1).

- Estimator: $\tilde{\theta}$
- Unknown parameter: $\theta$

By the Central Limit Theorem,

$$\frac{\tilde{\theta} - \theta}{\underbrace{g(\theta)}_{\text{sd}}/\sqrt{n}} \sim G(0,1) \quad \text{for large } n$$

Ex.

## Example: Binomial Experiment

- Consider a binomial experiment where $\hat{\theta} = \frac{y}{n}$ and $\tilde{\theta} = \frac{Y}{n}$
  - $E(\tilde{\theta}) = \theta$
  - $sd(\tilde{\theta}) = \sqrt{\frac{\theta(1-\theta)}{n}}$
- Show, using the Central Limit Theorem, how we can construct an asymptotic pivotal quantity $Q_n(\tilde{\theta}; \theta)$ that follows (approximately) a $G(0,1)$ distribution. What is $g(\theta)$ in your expression?

Using $\frac{\tilde{\theta} - \theta}{g(\theta)/\sqrt{n}} \sim G(0,1)$

## Choosing a Sample Size for a Binomial Experiment

- **Example**: show that choosing a sample size of $n \geq 1068$ will result in an approximately 95% confidence interval that is no wider than $2(0.03)$.