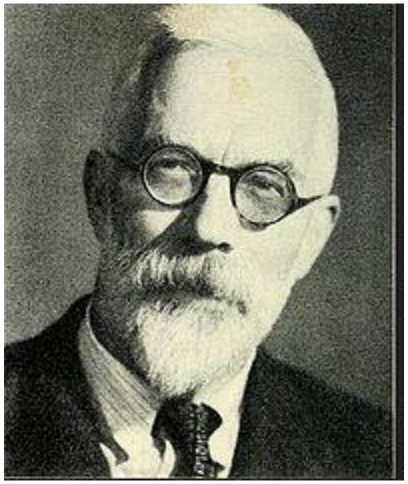
Confidence Intervals

1. Point and maximum likelihood estimation
2. Confidence intervals on the mean and variance
3. In-class exercise
4. Central limit theorem

Confidence Intervals

Point and Maximum Likelihood Estimation



Ronald Fisher
1956

Motivation

- Parameters of distribution functions can be estimated from samples
- Normal distribution
  - » True mean μ and variance σ^2
  - » Estimated mean $\bar{x}$ and variance s^2
- Because the number of samples is finite, these estimate will have uncertainty
- The objective is to quantify the uncertainty as:

$$CONF_{\gamma}\left\{\theta_1 \le \theta \le \theta_2\right\}$$

  - » θ is the parameter estimate
  - » θ_1 and θ_2 define the uncertainty range
  - » γ is the confidence level that θ is contained within the range
  - » The larger γ is chosen, the larger the parameter range

Random Sampling

- Random variable *X*
  - » Occur frequently in chemical engineering applications
  - » Molecular weight of a polymer product
  - » Film thickness of a solar cell
- Random sampling
  - » Obtain samples from a population with unknown statistics
  - » All outcomes must be equally likely to be sampled
  - » Meaningful statistics can be obtained from the samples
- Sample mean and variance

Sample mean $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ Sample variance $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$

Point Estimation

- Point estimates
  - » Estimates that approximate unknown parameter values of the population from which the samples were randomly selected are calculated from the samples
- Gaussian distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \hat{\mu} \quad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \hat{\sigma}^2$$

- Binomial distribution

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad \mu = np \quad \Rightarrow \quad \hat{p} = \frac{\bar{x}}{n} = \frac{\hat{\mu}}{n}$$

Maximum Likelihood Estimation

- Consider a random variable X with probability distribution function depending on a single parameter $f(x,\theta)$
- Collect n random samples $\{x_1, x_2, \ldots, x_n\}$
- Likelihood function: probability that a sample of size n will consist of precisely these values

$$l(x_1, \ldots x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$

- Select θ to maximize the likelihood:

$$\frac{\partial l}{\partial \theta} = 0 \quad \Rightarrow \quad \frac{\partial \ln l}{\partial \theta} = 0$$

Normal Distribution

- Multiple parameters: $f(x,\mu,\sigma)$

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \frac{\partial \ln l}{\partial \mu} = 0 \qquad \frac{\partial \ln l}{\partial \sigma} = 0$$

- Likelihood function:

$$l(x_1,\ldots x_n,\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2\right] \cdots \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-h}$$

$$h \equiv \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

Normal Distribution

- Logarithm of likelihood function:

$$\ln l = -n \ln \sqrt{2\pi} - n \ln \sigma - h$$

- Estimates:

$$\frac{\partial \ln l}{\partial \mu} = -\frac{\partial h}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu) = 0$$

$$\sum_{j=1}^{n} x_j - n\mu = 0 \qquad \hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} x_j = \bar{x}$$

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^{n} (x_j - \mu)^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \hat{\mu})^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 \neq s^2$$

UMASS.

Confidence Intervals

Confidence Intervals on the Mean and Variance

Jerzy Neyman
1937

Confidence Intervals

- Definition
  - » An interval in which the unknown parameter is contained with a certain probability γ: $\mathrm{CONF}_\gamma\{\theta_1 \le \theta \le \theta_2\}$
  - » γ = confidence level
  - » θ_1, θ_2 = confidence limits depending on γ
- *t*-distribution (governs confidence interval on mean)

$$F(z) = K_m \int_{-\infty}^{z} \left(1 + \frac{u^2}{m}\right)^{-(m+1)/2} du \quad F(\infty) = 1$$

  - » *m* = degrees of freedom
  - » Values tabularized (Table A9 in Appendix 5)
- Chi-square distribution (governs confidence interval on variance)

$$F(z) = C_m \int_{0}^{z} e^{-u/2} u^{(m-2)/2} du \quad z \ge 0$$

  - » Values tabularized (Table A10 in Appendix 5)

Confidence Intervals on the Mean

- Consider a normal distribution with unknown μ and σ^2
- Choose the confidence level γ
- Use the following equation to determine the value *c* from the *t*-distribution with $m = n$-1:

$$F(c) = \tfrac{1}{2}(1+\gamma)$$

- Compute the mean and variance of the sample $\{x_1, x_2, \ldots, x_n\}$
- Confidence interval:

$$\mathrm{CONF}_{\gamma}\ \{\bar{x} - k \le \mu \le \bar{x} + k\} \quad k = cs/\sqrt{n}$$

- The text also shows how to compute confidence intervals on the mean for known σ^2

Mean Confidence Interval Example 1

- Measurements of polymer molecular weight (scaled by 10^{-5})

$$\{1.25 \quad 1.36 \quad 1.22 \quad 1.19 \quad 1.33 \quad 1.12 \quad 1.27 \quad 1.27 \quad 1.31 \quad 1.26\}$$

- Confidence interval

$$\gamma = 0.95 \quad m = 10 - 1 = 9 \quad F(c) = 0.975 \quad c = 2.26$$

$$\bar{x} = 1.26 \quad s^2 = 0.0049$$

$$k = \frac{(2.26)\sqrt{0.0049}}{\sqrt{10}} = 0.050 \quad CONF_{0.95}\{1.21 \le \mu \le 1.31\}$$

Mean Confidence Interval Example 2

- Consider a set of measurements with sample mean $\bar{x} = 1$ and variance $s^2 = 0.1$
- Assuming that $\bar{x}$ and s^2 are unchanged by the addition of samples, determine the sample size n needed such that the mean has the following confidence limits:

$$\mathrm{CONF}_{0.95}\{0.9 \le \mu \le 1.1\} \quad \Rightarrow \quad k = 0.1$$

- Trial-and-error solution using Table A9 shows that at least 7 samples are needed:

n	m	c	k
6	5	2.57	0.105
7	6	2.45	0.093
8	7	2.36	0.083

Confidence Intervals on the Variance

- Consider a Gaussian distribution with unknown μ and σ^2
- Choose the confidence level γ
- Use the following equations to determine values c_1 and c_2 from the chi-squared distribution with $m = n$-1:

$$F(c_1) = \tfrac{1}{2}(1-\gamma) \quad F(c_2) = \tfrac{1}{2}(1+\gamma)$$

- Compute the variance of the sample $\{x_1, x_2, \ldots, x_n\}$
- Confidence interval:

$$\mathrm{CONF}_\gamma \{k_2 \le \sigma^2 \le k_1\} \quad k_1 = (n-1)s^2 / c_1 \quad k_2 = (n-1)s^2 / c_2$$

Variance Confidence Interval Example 1

- Measurements of polymer molecular weight (scaled by 10^{-5})

$$\{1.25 \quad 1.36 \quad 1.22 \quad 1.19 \quad 1.33 \quad 1.12 \quad 1.27 \quad 1.27 \quad 1.31 \quad 1.26\}$$

- Confidence interval

$$\gamma = 0.95 \quad m = 9 \quad F(c_1) = 0.025 \quad c_1 = 2.70 \quad F(c_2) = 0.975 \quad c_2 = 19.02$$

$$s^2 = 0.0049 \quad k_1 = \frac{(9)(0.0049)}{2.70} = 0.0163 \quad k_2 = \frac{(9)(0.0049)}{19.02} = 0.0023$$

$$CONF_{0.95}\{0.002 \le \sigma^2 \le 0.016\}$$

Variance Confidence Interval Example 2

- Consider a set of measurements with sample mean $\bar{x} = 1$ and variance $s^2 = 0.1$
- Assuming that $\bar{x}$ and s^2 are unchanged by the addition of samples, determine the sample size n needed such that the variance has the following confidence limits:

$$\mathrm{CONF}_{0.95}\{\sigma^2 \le 0.11\} \quad \Rightarrow \quad k_1 = 0.11$$

- Trial-and-error solution using Table A10 shows that more than 100 samples are needed:

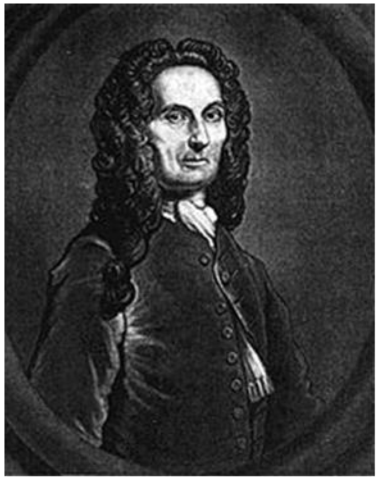
n	**m**	**c_2**	**k_1**
10	9	2.70	0.333
101	100	74.2	0.135

Confidence Intervals

In-class Exercise

Confidence Intervals

Central Limit Theorem



Abraham de Moivre
1733

Central Limit Theorem

- The confidence limit methods were developed only for the normal distribution
- The central limit theorem justifies the application of these methods to other distributions for large sample sizes
- Let $\{X_1,\ldots,X_n\}$ be independent random variables, each with the same mean μ and variance σ^2. Then:
  - » $Y_n = X_1+\cdots+X_n$ has the mean $n\mu$ and variance $n\sigma^2$
  - » If $\{X_1,\ldots,X_n\}$ are also normal variables, then Y_n is a normal random variable

Central Limit Theorem

- Consider following random variable Z_n:

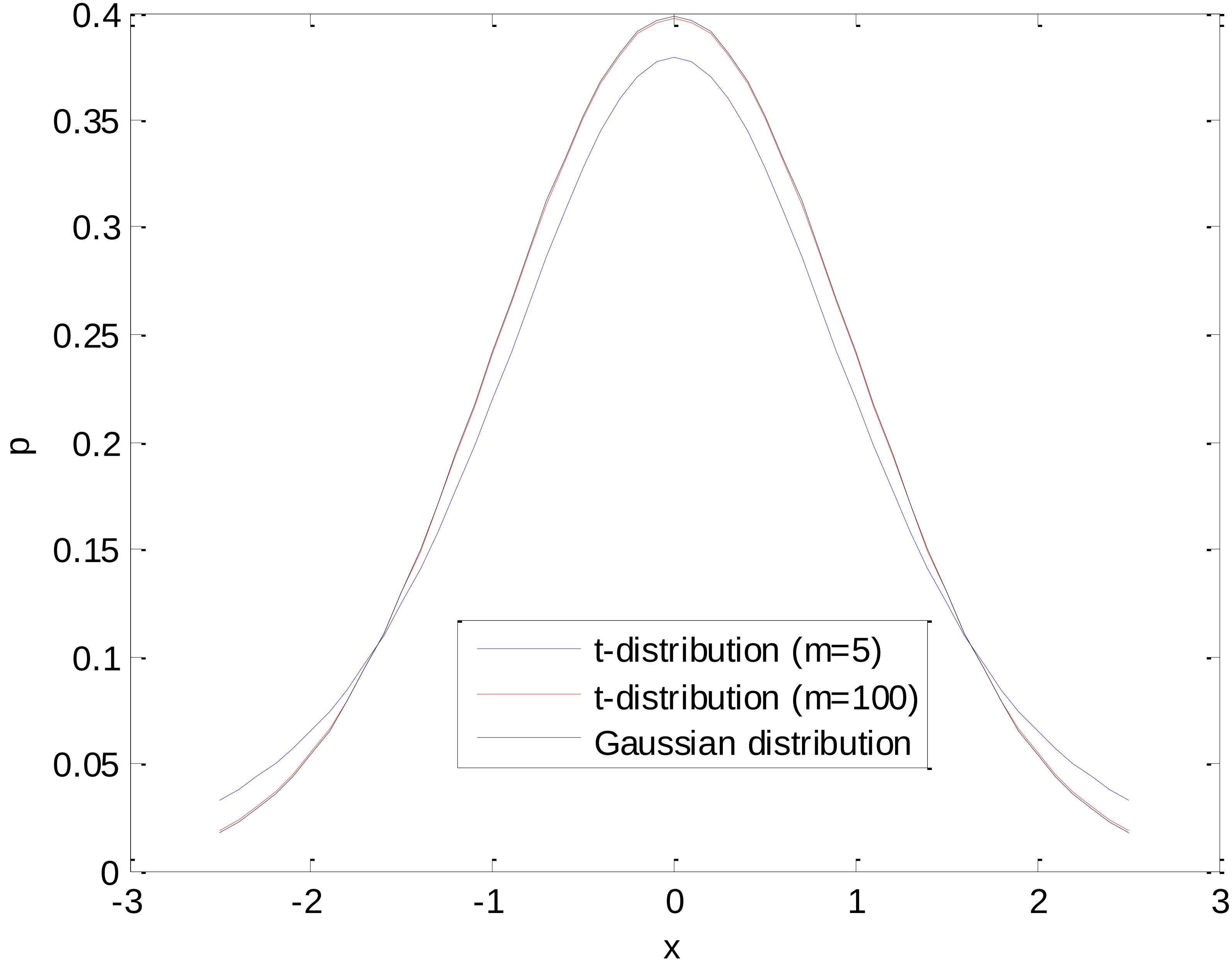
$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

- Central limit theorem:
  - » Z_n is asymptotically normal with zero mean and unity variance in the sense that its distribution function $F_n(x)$ satisfies:

$$\lim_{n\to\infty} F_n(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

- Implication: can determine confidence intervals for non-normal distributions using previous methods if sufficiently large sample sizes are used ($n > 20$ for mean; $n > 50$ for variance)

Testing the Central Limit Theorem



Testing the Central Limit Theorem