EC381/MN308
Probability and Some Statistics

Yannis Paschalidis

yannisp@bu.edu, http://ionia.bu.edu/

Dept. of Manufacturing Engineering

Dept. of Electrical and Computer Engineering

Center for Information and Systems Engineering

Lecture 20 - Outline

1. **Sample Mean and Sample Variance**
2. **Confidence Intervals (partly in the textbook)**
3. **Control Charts (not in the textbook)**
4. **Sampling distributions (not in the textbook)**
5. **More general confidence intervals (not in the textbook)**

Sample Mean

- An experiment characterized by an RV X with mean μ_X and variance σ_X^2. Consider i.i.d. samples $X_1, X_2, \ldots, X_n$
- Estimator of the mean: The sample mean

$$M_n(X) = \frac{\sum_{i=1}^n X_i}{n}$$

- Recall

$$E[M_n(X)] = \mu_X \qquad \textbf{(unbiased)}$$

$$Var(M_n(X)) = \frac{\sigma_X^2}{n}$$

$$e_n = E[(M_n(X) - \mu_X)^2] = Var(M_n(X)) = \frac{\sigma_X^2}{n}$$

Sample Variance (known mean)

- Estimator of the variance: The sample variance

$$W_n(X) = \frac{\sum_{i=1}^n (X_i - \mu_X)^2}{n}$$

- Note

$$E[W_n(X)] = \frac{n\sigma_X^2}{n} = \sigma_X^2 \qquad \textbf{(unbiased)}$$

Sample Variance (unknown mean)

$$S_n(X) = \frac{\sum_{i=1}^n (X_i - M_n(X))^2}{n-1}$$

- We have

$$E[S_n(X)] = \sigma_X^2 \qquad \textbf{(unbiased)}$$

Proof:

$$\begin{aligned}
E[S_n(X)] &= \frac{1}{n-1} E[\sum_i (X_i - \sum_j X_j/n)^2] \\
&= \frac{1}{n-1} E\left[\sum_i \left(X_i^2 + \frac{(\sum_j X_j)^2}{n^2} - 2\frac{X_i}{n}\sum_j X_j\right)\right] \\
&= \frac{1}{n-1} E\left[\sum_i X_i^2 + \frac{n}{n^2}(\sum_j X_j)^2 - \frac{2}{n}(\sum_i X_i)(\sum_j X_j)\right] \\
&= \frac{1}{n-1} E\left[\sum_i X_i^2 - \frac{1}{n}\sum_i\sum_j X_i X_j\right] \\
&= \frac{1}{n-1} nE[X^2] - \frac{1}{n(n-1)} nE[X^2] - \frac{1}{n(n-1)} n(n-1)\mu_X^2 \\
&= \sigma_X^2
\end{aligned}$$

Confidence Intervals

- Suppose we are interested in estimating a model parameter, say r, e.g., $M_n(X)$ is an estimate of E[X]
- Find random variables A, B so that

$$P[\underset{\text{Lower confidence limit}}{A} \le r \le \underset{\text{Upper confidence limit}}{B}] \ge \underbrace{1-\alpha}_{\text{Confidence coefficient}}$$

- B-A is called the confidence interval
  - e.g., r∈ [A,B] with a 95% confidence
  - better to have small B-A and high 1-α

Confidence interval for the mean (known variance)

$$P[M_n(X) - c < \mu_X < M_n(X) + c] \geq 1 - \frac{\sigma_X^2}{nc^2}$$

Proof:

$$\begin{aligned}
&P[M_n(X) - c < \mu_X < M_n(X) + c] \\
=&P[c > M_n(X) - \mu_X > -c] \\
=&P[|M_n(X) - \mu_X| < c] \\
=&1 - P[|M_n(X) - \mu_X| \geq c] \qquad \text{(Chebyshev)} \\
\geq&1 - \frac{\sigma_X^2}{nc^2}
\end{aligned}$$

Example 7.9 (modified)

Measure (or simulate) a quantity of interest b. Measurements (i.i.d.):

$X_i = b+W_i$

where W_i is zero mean noise with variance 1. How many measurements (or simulation runs) do I need to get a confidence interval for b of length 0.2 with 99% confidence level?

$$E[X_i] = b, \qquad Var(X_i) = Var(W_i) = 1$$

$$\begin{aligned}
&P[M_n(X) - 0.1 < b < M_n(X) + 0.1] \\
&\geq 1 - \frac{1}{n(0.1)^2} \\
&= 1 - \frac{100}{n} \geq 0.99 \\
\Rightarrow &\frac{100}{n} \leq 0.01 \\
\Rightarrow &n \geq 10,000
\end{aligned}$$

Confidence interval for the mean (known variance)

- So far we used Chebyshev's inequality (which can be loose) to get a bound on the confidence level.
- Can we do better, especially if n is large?
- An experiment characterized by an RV X with unknown mean μ_X and known variance σ_X^2. Collect i.i.d. samples $X_1, X_2, \ldots, X_n$
- Estimator: the sample mean

$$M_n(X) = \frac{\sum_{i=1}^n X_i}{n} \qquad E[M_n(X)] = \mu_X \qquad Var(M_n(X)) = \frac{\sigma_X^2}{n}$$

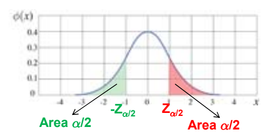
Confidence interval for the mean (cont.)

- By the Central Limit Theorem, as $n \to \infty$, we have

$$Z_n = \frac{M_n(X) - \mu_X}{\sigma_X/\sqrt{n}} \xrightarrow{D} Z = N(0,1)$$

- Let $Z_{\alpha/2}$ such that

$$P[Z > Z_{\alpha/2}] = Q\left(Z_{\alpha/2}\right) = \frac{\alpha}{2}$$



- It follows

$$P[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}] = 1 - \alpha$$

$$P[-Z_{\alpha/2} \leq \frac{M_n(X) - \mu_x}{\sigma_X/\sqrt{n}} \leq Z_{\alpha/2}] = 1 - \alpha, \quad \text{for large } n$$

- Confidence Interval (shrinks as sample size increases)

$$\boxed{M_n(X) - \frac{\sigma_X}{\sqrt{n}} Z_{\alpha/2} \leq \mu_X \leq M_n(X) + \frac{\sigma_X}{\sqrt{n}} Z_{\alpha/2}}$$

Example

Measure (or simulate) response time of a service system.

Measurements: 41.6, 41.48, 42.34, 41.95, 41.86, 42.18, 41.72, 42.26, 41.81, 42.04 (n=10)

Sample mean: $M_{10}(X)$=41.924

Suppose we know: $\sigma_X = 0.1$

Find 95% confidence interval:

$$1 - \alpha = 0.95 \Rightarrow \frac{\alpha}{2} = 0.025$$

$$Q(Z_{\alpha/2}) = \frac{\alpha}{2} \Rightarrow \Phi(Z_{\alpha/2}) = 1 - \frac{\alpha}{2} = 0.975 \Rightarrow Z_{\alpha/2} = 1.96$$

$$41.924 - \frac{0.1}{\sqrt{10}} 1.96 \leq \mu_X \leq 41.924 + \frac{0.1}{\sqrt{10}} 1.96$$

$$41.862 \leq \mu_X \leq 41.986$$

If we are interested in 99.73% confidence then $Z_{\alpha/2}$=3

Example 7.9 (revisited)

Measure (or simulate) a quantity of interest b. Measurements (i.i.d.):

$X_i = b+W_i$

where W_i is zero mean noise with variance 1. Find a 99% confidence interval when n=10,000.

$$E[X_i] = b, \qquad Var(X_i) = Var(W_i) = 1$$

$$1 - \alpha = 0.99 \Rightarrow \frac{\alpha}{2} = 0.005$$

$$Q(Z_{\alpha/2}) = \frac{\alpha}{2} \Rightarrow \Phi(Z_{\alpha/2}) = 0.995 \Rightarrow Z_{\alpha/2} = 2.58$$

$$M_n(X) - \frac{1}{\sqrt{10^4}} 2.58 \leq b \leq M_n(X) + \frac{1}{\sqrt{10^4}} 2.58$$

$$M_n(X) - 0.0258 \leq b \leq M_n(X) + 0.0258$$

Much tighter than the requirement of confidence interval length = 0.2.

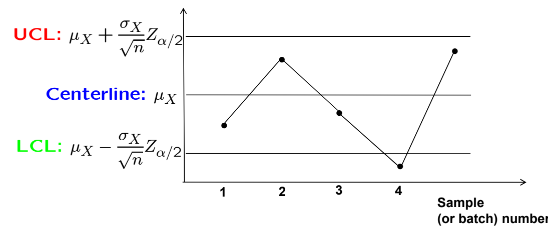
An Application in Quality Control: Control Charts

- Suppose we monitor some manufacturing process, or some quantity of interest (e.g., dimension of a part, temperature, etc.).
- A sample (or batch) consists of $X_1,\ldots,X_n$ i.i.d. measurements with mean μ_X and variance σ_X^2
- Interested in determining if process is “OK”.
- Compute the sample mean $M_n(X)$. For large n CLT applies

$$P[-Z_{\alpha/2} \le \frac{M_n(X)-\mu_x}{\sigma_X/\sqrt{n}} \le Z_{\alpha/2}] = 1-\alpha$$

$$\Rightarrow P\left[\underbrace{\mu_X - \frac{\sigma_X}{\sqrt{n}} Z_{\alpha/2}}_{\text{Lower Control Limit (LCL)}} \le M_n(X) \le \underbrace{\mu_X + \frac{\sigma_X}{\sqrt{n}} Z_{\alpha/2}}_{\text{Upper Control Limit (UCL)}}\right] = 1-\alpha$$

Control Charts (cont.)



- Each sample (batch) has size n. Measure $M_n(X)$ and if within limits we say the process is “in control”

95% confidence: $Z_{\alpha/2} = 1.96 \approx 2$

99.73% confidence: $Z_{\alpha/2} = 3$ **(6σ interval)**

Control Charts (cont.)

- If we don’t know the mean:
  - Estimate from past samples (batches) that we know are in control
  - Let $M_{n,1}(X),\ldots, M_{n,k}(X)$ sample means from batches 1,…,k, each of size n

$$\hat{\mu}_{n,k}(X) = \frac{\sum_{i=1}^k M_{n,i}(X)}{k} = \frac{\sum_{i=1}^k (\sum_{j=1}^n X_{i,j})/n}{k}$$

- If we don’t know the variance:

$$\hat{S}_{n,k}(X) = \frac{\sum_{i=1}^k \sum_{j=1}^n (X_{i,j} - \hat{\mu}_{n,k}(X))^2}{nk-1}$$

$$\hat{\sigma}_{nk}(X) = \sqrt{\hat{S}_{n,k}(X)}$$

Control Charts (cont.)

- If batch size n is small and CLT does not provide a good approximation:
  - In practice we use

UCL: $\hat{\mu}_{n,k}(X) + 3\frac{\hat{\sigma}_{n,k}(X)}{\sqrt{n}}$

LCL: $\hat{\mu}_{n,k}(X) - 3\frac{\hat{\sigma}_{n,k}(X)}{\sqrt{n}}$

Control Charts for Attributes:

- Suppose we are interested in the fraction of defects produced by a system
- Sample (or batch) $X_1,\ldots,X_n$, i.i.d., of size n.
- $X_{i,j}$ =1{defect} is Bernoulli with unknown p.

$$D_n = \sum_{j=1}^n X_j$$

D_n: # of defects in batch, binomial distribution

$$\hat{p}_n = \frac{D_n}{n}$$

$$Var(\hat{p}_n) = \frac{np(1-p)}{n^2} \Rightarrow \hat{\sigma}^2_{\hat{p}_n} = \frac{\hat{p}_n(1-\hat{p}_n)}{n}$$

Control Charts for Attributes (cont.)

- From k samples (or batches) $X_{i,1},\ldots,X_{i,n}$, i.i.d., of size n, that we know are “in control” we estimate

Centerline: $\bar{p}_{n,k} = \frac{\sum_{i=1}^k \sum_{j=1}^n X_{i,j}}{nk}$

UCL $= \bar{p}_{n,k} + 3\sqrt{\frac{\bar{p}_{n,k}(1-\bar{p}_{n,k})}{n}}$

LCL $= \bar{p}_{n,k} - 3\sqrt{\frac{\bar{p}_{n,k}(1-\bar{p}_{n,k})}{n}}$

Control Charts: Use of distributional information

- Suppose we are interested in # of arrivals per time interval to a service system (e.g., communication switch, Sonsie's on a Saturday night, a manufacturing system, etc.)
- Assume a Poisson arrival process with arrival rate λ (arrivals per minute) ➔ # of arrivals A in interval of length T is Poisson RV with parameter $\mu=\lambda$ T

$$P_A(a) = \frac{\mu^a e^{-\mu}}{a!}, \qquad E[A] = \mu, \qquad Var(A) = \mu$$

- Observe intervals with # of arrivals $A_1,\ldots,A_k$

Centerline: $\bar{A}_k = \frac{\sum_{i=1}^k A_i}{k}$

UCL $= \bar{A}_k + 3\sqrt{\bar{A}_k}$ **LCL** $= \bar{A}_k - 3\sqrt{\bar{A}_k}$

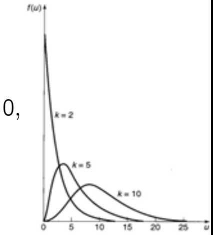
Sampling Distributions: Chi-Square

- $Z_1,\ldots,Z_k$ ~ N(0,1) i.i.d. Chi-Square distribution with k degrees of freedom

$$\chi_k^2 = Z_1^2 + \cdots + Z_k^2$$

$$f_{\chi_k^2}(u) = \frac{1}{2^{k/2}\Gamma(k/2)} u^{k/2-1} e^{-u/2}, \; u > 0,$$

$$E[\chi_k^2] = k, \qquad Var(\chi_k^2) = 2k$$



- As k➔∞, χ_k^2 approaches the Gaussian by CLT

$\chi^2_{\alpha,k}$: percentage point of χ_k^2 such that $P[\chi_k^2 \geq \chi^2_{\alpha,k}] = \alpha$

Chi-Square (cont.)

- Additivity Property: Let $\chi_1^2,\ldots,\chi_p^2$ i.i.d. chi-square RVs with $k_1,\ldots,k_p$ degrees of freedom, respectively. Then Y= $\chi_1^2+\ldots+\chi_p^2$ follows chi-square with $k_1+\ldots+k_p$ degrees of freedom

Proof:

Each chi-square can be written as a sum of squares of Gaussian.

- Important Property: Let $X_1, \ldots X_n$ i.i.d. Gaussian. Then

$$\frac{(n-1)S_n(X)}{\sigma_X^2}$$

follows χ_{n-1}^2 where $S_n(X)$ is the sample variance

$$S_n(X) = \frac{\sum_{i=1}^n (X_i - M_n(X))^2}{n-1}$$

Sampling Distributions: (student) t-distribution

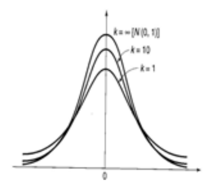
- Z~ N(0,1), V~χ_k^2 and Z, V independent. Then

$T = \frac{Z}{\sqrt{V/k}} \sim t_k$: t-distribution with k degrees of freedom

$$f_{t_k}(t) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\,\Gamma(k/2)} \frac{1}{(t^2/k+1)^{(k+1)/2}}$$

$$E[t_k] = 0,$$

$$Var(t_k) = k/(k-2) \; k > 2$$



- As k➔∞, t_k approaches the Gaussian

$t_{\alpha,k}$: percentage point of t_k such that $P[t_k \geq t_{\alpha,k}] = \alpha$

t distribution (cont.)

- Important Property: Let $X_1, \ldots X_n$ i.i.d. Gaussian. Then

$\frac{M_n(X) - \mu_X}{\sqrt{S_n(X)/\sqrt{n}}}$ follows t_{n-1} where $S_n(X)$ is the sample variance

$$S_n(X) = \frac{\sum_{i=1}^n (X_i - M_n(X))^2}{n-1}$$

Proof:

$$\frac{M_n(X) - \mu_X}{\sqrt{S_n(X)/\sqrt{n}}} = \frac{\frac{M_n(X)-\mu_X}{\sigma_X/\sqrt{n}}}{\sqrt{S_n(X)/\sigma_X^2}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

Confidence interval for the variance

- Let i.i.d. samples $X_1, X_2, \ldots, X_n$ from Gaussian (each X_i can be a large sum satisfying CLT)
- We know $\frac{(n-1)S_n(X)}{\sigma_X^2}$ ~ χ^2_{n-1}
- By the definition of percentage points (χ^2 is not symmetric)

$$P\left[\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S_n(X)}{\sigma_X^2} \leq \chi^2_{\alpha/2,n-1}\right] = 1-\alpha$$

- Thus, a 100(1-α) % confidence interval for the variance is

$$\boxed{\frac{(n-1)S_n(X)}{\chi^2_{\alpha/2,n-1}} \leq \sigma_X^2 \leq \frac{(n-1)S_n(X)}{\chi^2_{1-\alpha/2,n-1}}}$$

Confidence interval for mean (unknown variance)

- Let i.i.d. samples $X_1, X_2, \ldots, X_n$ from Gaussian (each X_i can be a large sum satisfying CLT)
- We know $\frac{M_n(X) - \mu_X}{\sqrt{S_n(X)}/\sqrt{n}} \sim t_{n-1}$
- By the definition of percentage points (t is symmetric)

$$P\left[-t_{\alpha/2,n-1} \le \frac{M_n(X) - \mu_X}{\sqrt{S_n(X)}/\sqrt{n}} \le t_{\alpha/2,n-1}\right] = 1-\alpha$$

- Thus, a 100(1-α) % confidence interval for the mean is (as n→∞ it becomes the same as one with known variance)

$$\boxed{M_n(X) - t_{\alpha/2,n-1}\frac{\sqrt{S_n(X)}}{\sqrt{n}} \le \mu_X \le M_n(X) + t_{\alpha/2,n-1}\frac{\sqrt{S_n(X)}}{\sqrt{n}}}$$