Comparing Systems Using Sample Data

Dr. John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@cs.rice.edu

Goals for Today

Understand

- **Population and samples**
- **Confidence intervals**
- **How to compute confidence intervals for**
  - —sample mean
  - —proportions
- **How to test a hypothesis using a test for zero mean**
- **How to compare two alternatives**
- **How to determine minimum sample size for estimating**
  - —sample mean with a given accuracy
  - —proportion with a given accuracy

Sample vs. Population

- **Suppose we generate a set S containing several million random numbers. We will call this set the *population*.**
  - —denote population mean with μ
  - —denote population std deviation with σ
- **Draw a sample of *n* numbers $\{x_1, x_2, \ldots, x_n\}$ from S**
  - —denote sample mean with $\bar{x}$
  - —denote the std deviation of the sample with *s*
- **No guarantee that $\bar{x} = \mu \wedge s = \sigma$**
- **($\bar{x}$,s) of sample are *estimates* of the population parameters (μ, σ)**
- **Conventions**
  - —population characteristics: *parameters* (μ, σ) (Greek alphabet)
  - —sample estimates: *statistics* ($\bar{x}$,s) (Roman alphabet)

Confidence Interval for the Mean

- ***k*** **samples of a population may yield** ***k*** **different sample means**
- **No sample or finite set of samples will necessarily give a perfect estimate for the population mean μ**
- **Instead, we use probability bounds for an estimate of μ, the population mean**

$$P(c_1 \le \mu \le c_2) = 1 - \alpha$$

- **Confidence interval (c_1,c_2)**
- **Significance level**
- **Confidence coefficient**
- **Confidence level (a percentage): 100(1- α)**

Understanding Confidence Intervals

- **Why use them?**
  - —provide a way to decide if measurements are meaningful
  - —characterize potential error in sample mean
  - —enable comparisons in the presence of experimental error
- **Understand their limitations!**
  - — at 95% confidence, confidence intervals for 5% of sample means *do not* include the population mean μ

Computing (c_1,c_2) for Population Mean μ

The hard way

To compute a 90% confidence interval for a population mean μ

- **Take *k* samples of the population (each sample is a set)**
- **Compute the set of sample means (one for each sample)**
- **Sort the set of sample means**
- **Select the $[1 + .05(k-1)]^{th}$ element as c_1**
- **Select the $[1 + .95(k-1)]^{th}$ element as c_2**
- **90% confidence interval for μ is (c_1, c_2)**

90% = 100(1- α)

.05 = $\alpha/2$

.95 = 1- $\alpha/2$

The Central Limit Theorem

- **If observations $\{x_1, x_2, \ldots, x_n\}$ are**
  - **—independent**
  - **—from the same population**
  - **—the population has mean μ**
  - **—the population has std deviation σ**
- **Then sample mean $\bar{x}$ for large samples is *approximately* normally distributed**

$$\bar{x} \sim N(\mu, \sigma / \sqrt{n})$$

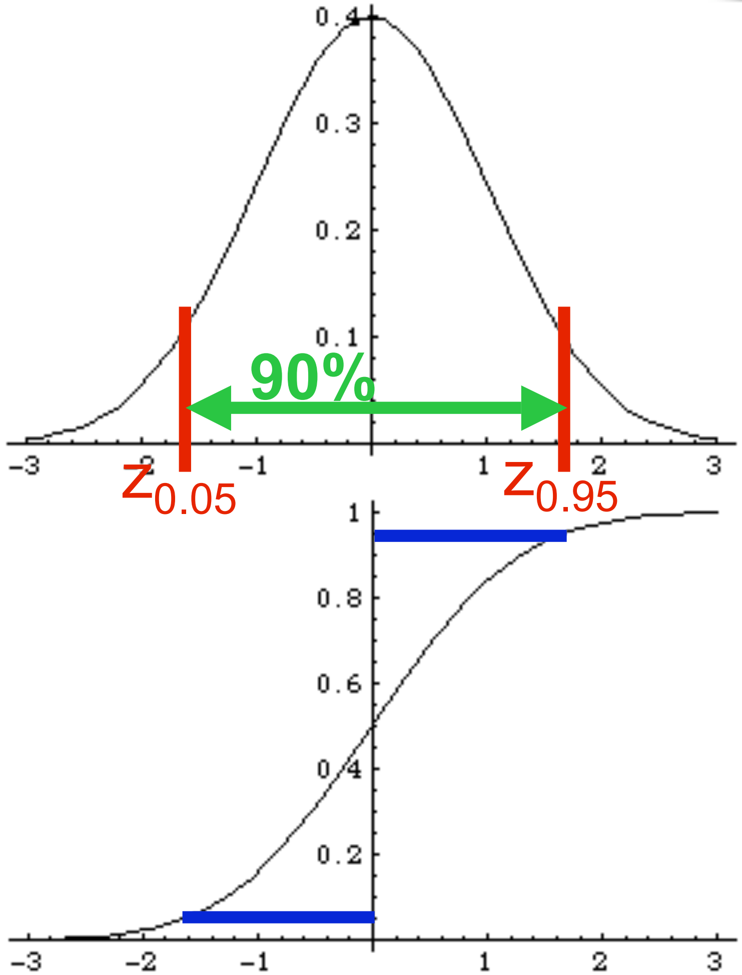
- **Std error = std deviation of sample mean**
- **If population std deviation is σ, std error is $\sigma / \sqrt{n}$**

Confidence Interval of a Normal Distribution

Example: 90% confidence interval. $\alpha = .10$

PDF of N(0,1)

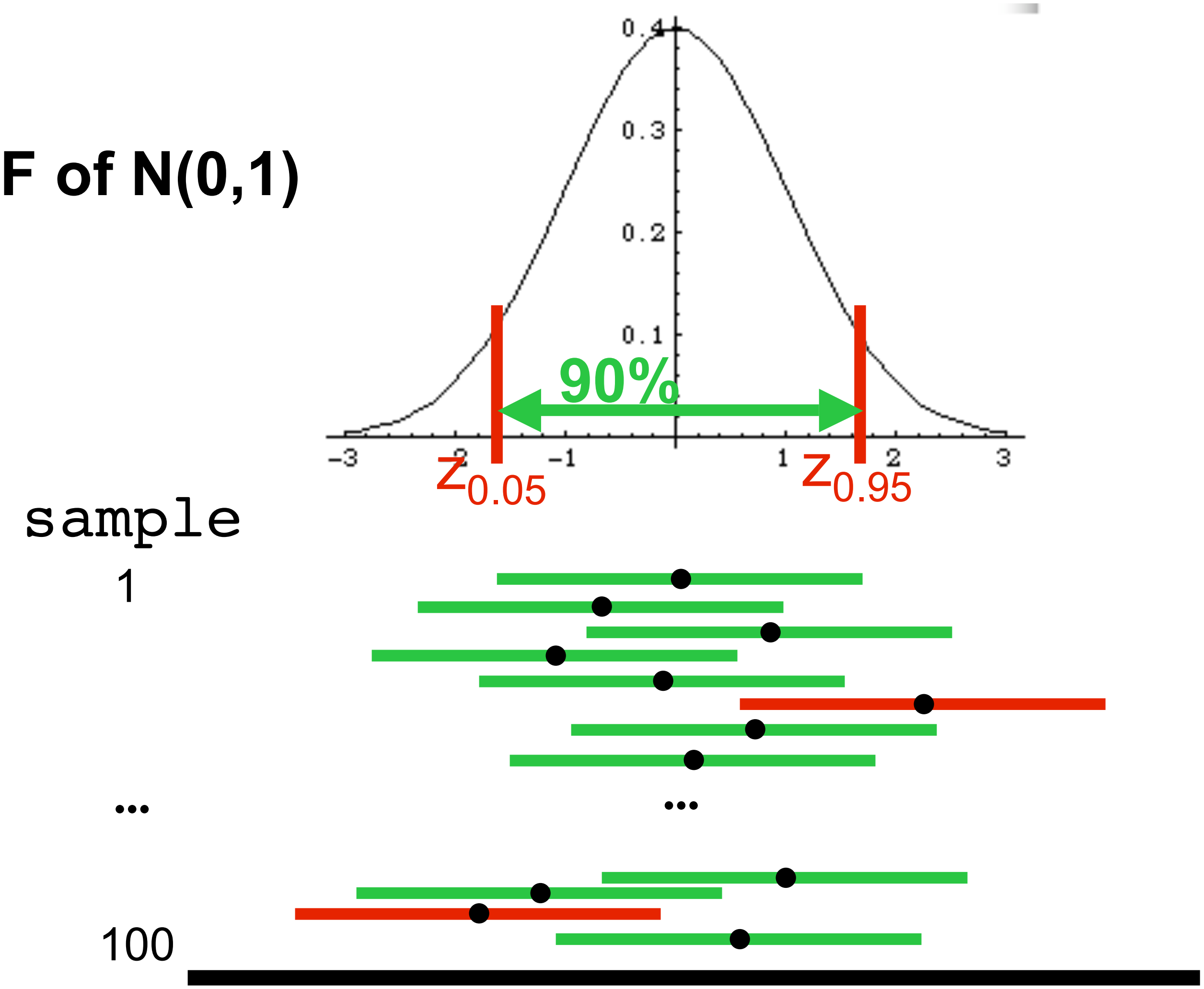
CDF of N(0,1)



Meaning of a Confidence Interval

Example: 90% confidence interval. $\alpha = .10$

PDF of N(0,1)



total yes = 90 = 100(1-.10)

total no = 10 = 100 (.10)

Computing (c_1,c_2) for a Population Mean μ

The easy way (for a *large* sample, n > 30)

By the central limit theorem, a 100(1- α)% confidence interval for μ

$$(\bar{x} - z_{1-\alpha/2}\boxed{s/\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\boxed{s/\sqrt{n}})$$

scaled by std error

Where

- $\bar{x}$ **is the sample mean**
- **s is the sample std deviation**
- **n is the sample size**
- **α is the significance level, 100(1- α)% is the confidence level**
- $z_{1-\alpha/2}$ **is the (1- α/2) quantile of the unit normal variate**

Confidence Interval Example

- **Given a (large) sample with the following characteristics**
  - **32 elements (n = 32)**
  - **sample mean $\bar{x}$ = 3.90**
  - **sample std deviation s = .71**
- **A 90% confidence interval for the mean can be computed as**

$$(\bar{x} - z_{1-\alpha/2} s/\sqrt{n}, \bar{x} + z_{1-\alpha/2} s/\sqrt{n})$$

$$(\bar{x} - (z_{0.95}) s/\sqrt{n}, \bar{x} + (z_{0.95}) s/\sqrt{n})$$

Recall $z_{1-\alpha/2}$ is approximately $4.91[(1-\alpha/2)^{0.14} - (\alpha/2)^{0.14}]$

$$(3.90 - (1.645)(0.71)/\sqrt{32}, 3.90 + (1.645)(0.71)/\sqrt{32})$$

$$= (3.69, 4.11)$$

Computing (c_1,c_2) for Population Mean μ

The easy way (for a *small* sample, n ≤ 30)

- **For smaller samples, confidence intervals can only be constructed if samples come from a normally distributed population**
- **Ratio of $(\bar{x} - \mu)/(s/\sqrt{n})$ follows a *t*(n-1) distribution**
- **For small samples, 100(1- α)% confidence interval for μ is**

$$(\bar{x} - t_{[1-\alpha/2;n-1]} s/\sqrt{n}, \bar{x} + t_{[1-\alpha/2;n-1]} s/\sqrt{n})$$

Where

$\bar{x}$ **is the sample mean**

s is the sample std deviation

n is the sample size

α **is the significance level, 100(1- α)% is the confidence level**

$t_{[1-\alpha/2;\, n-1]}$ **: (1- α/2) quantile of *t* distribution with n-1 degrees of freedom**

Confidence Interval of a t(n-1) Distribution

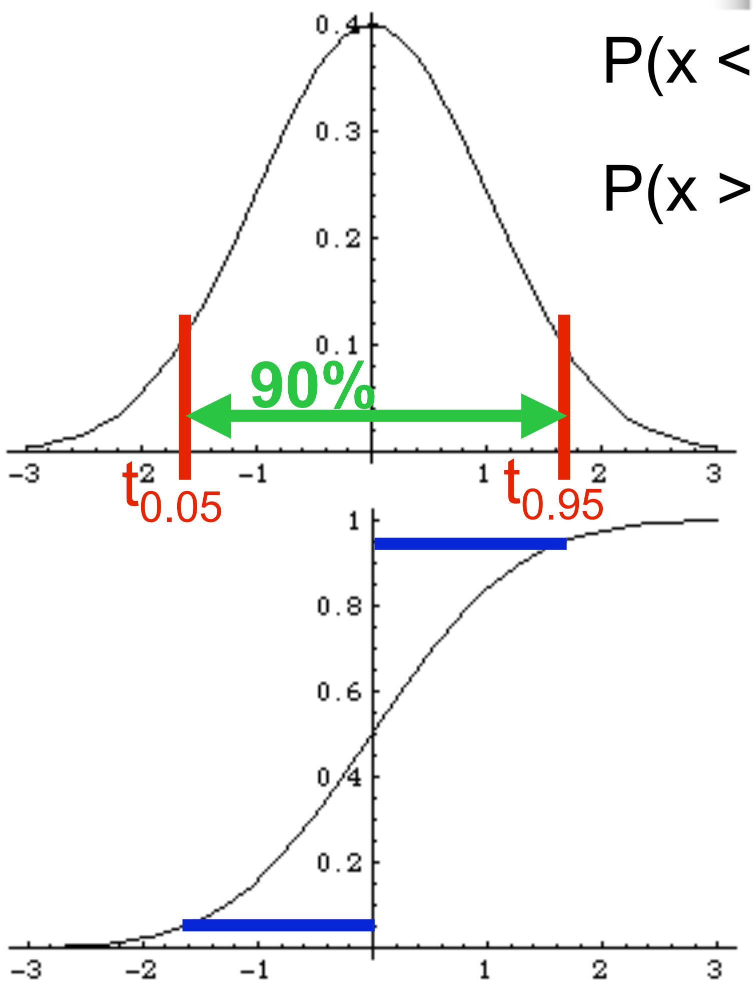
Example: 90% confidence interval

$P(x < -t_{[1-.05;\, n-1]}) = \alpha/2$

$P(x > t_{[1-.05;\, n-1]}) = \alpha/2$

PDF of t(n-1)

CDF of t(n-1)



Confidence Interval Example

- Given a (small) sample with the following characteristics
  - —modeling error shown to be normally distributed from quantile/quantile plot {-.04,-.19,.14,-.09,-.14,.19,.04,.09}
  - —8 elements (n = 8)
  - —sample mean $\bar{x}$ = 0
  - —sample std deviation s = .138
- A 90% confidence interval for the mean can be computed as

$$(\bar{x} - t_{[1-\alpha/2;n-1]} s/\sqrt{n}, \bar{x} + t_{[1-\alpha/2;n-1]} s/\sqrt{n})$$

$$(0 - (t_{[0.95;7]})(.138)/\sqrt{8}, 0 + (t_{[0.95;7]})(.138)/\sqrt{8})$$

Look up $t_{[1-.05;\ n-1]}$ in Jain Table A.4 = 1.895

$$(0 - (1.895)(.138)/\sqrt{8}, 0 + (1.895)(.138)/\sqrt{8})$$

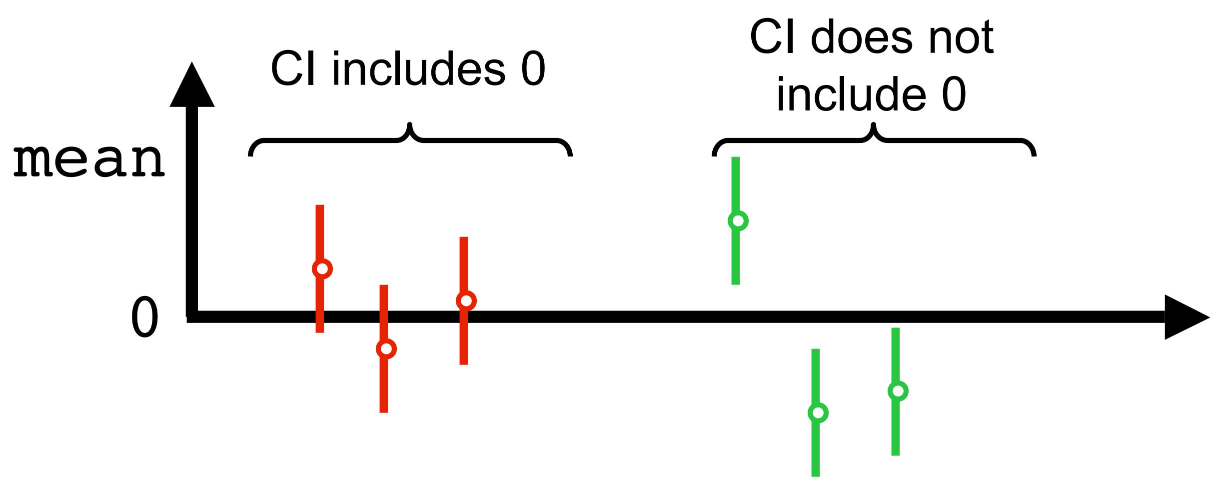
= (-.0926,+.0926)

Small vs. Large Samples

- **Why the difference when computing confidence for small vs. large samples?**
- **As n increases, t-distribution approaches normal distribution**

Testing for a Zero Mean

- **Is a measured value significantly different from zero?**
  - **—common use of confidence intervals**
- **When comparing random measurement with zero, must do so probabilistically**
- **If value different from zero with probability 100(1-α)%, then value is *significantly different* from zero**



Example: Testing for a Zero Mean

- **Difference in running time of two sorting algorithms A and B was measured on several different input sequences**
- **Differences are {1.5, 2.6, -1.8, 1.3, -.5, 1.7, 2.4}**
- **Can we say with 99% confidence that 1 algorithm is superior?**
- **Example properties**
  - —n = 7, $\bar{x}$ = 1.03, std deviation = 1.6
  - —α = .01, $\alpha/2$ = .005
  - —confidence interval $(1.03 - t_{[1-.005;6]} * 1.60/\sqrt{7}, 1.03 + t_{[1-.005;6]} * 1.60/\sqrt{7})$

Look up $t_{[1-.005;\, 6]} = t_{[.995;\, 6]}$ in Jain Table A.4 = 3.707

$$(1.03 - (3.707) * 1.60/\sqrt{7}, 1.03 + (3.707) * 1.60/\sqrt{7})$$

$$= (-1.21, 3.27)$$

Confidence interval includes 0; thus, cannot say with 99% confidence that the mean difference between A & B is significantly different from 0

Comparing Two Alternatives

- **Most performance analysis projects require determining the best of two or more systems**
- **Here we compare 2 systems with *very similar* workloads**
  - **—more than 2 systems or substantially different workloads**
    - **– use techniques for experimental design (covered later in course)**

Paired Observations

- **Conduct n experiments on each of 2 systems**
  - —system a: $\{a_1, a_2, \ldots, a_n\}$
  - —system b: $\{b_1, b_2, \ldots, b_n\}$
- **If one-one correspondence between tests on both systems**
  - —observations are said to be "paired"
- **Treat the samples for 2 systems as one sample of n pairs**
- **For each pair, compute difference in performance**
  - —$\{a_1\text{-}b_1, a_2\text{-}b_2, \ldots, a_n\text{-}b_n\}$
- **Construct a confidence interval for the mean difference**
- **Is the confidence interval includes 0, systems not significantly different**

Unpaired Observations (*t*-test)

- **Two samples, one size n_a, the other size n_b**
- **Compute mean of each sample:** $\bar{x}_a$, $\bar{x}_b$
- **Compute std deviation of each sample: s_a, s_b**
- **Compute mean difference** $\bar{x}_a - \bar{x}_b$
- **Compute std deviation of mean difference** $\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$
- **Effective number of degrees of freedom**

$$\nu = \frac{(s_a^2/n_a + s_b^2/n_b)^2}{\frac{1}{n_a+1}\left(\frac{s_a^2}{n_a}\right)^2 + \frac{1}{n_b+1}\left(\frac{s_b^2}{n_b}\right)^2} - 2$$

- **Confidence interval for mean difference**

$$(\bar{x}_a - \bar{x}_b) \pm t_{[1-\alpha/2;\nu]} s$$

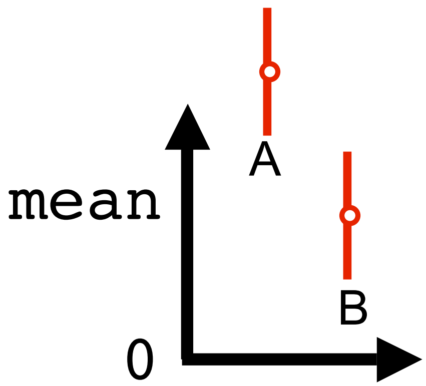
Notes on Unpaired Observations

- **Preceding slide made following assumptions**
  - —two samples of unequal size
  - —standard deviations not assumed equal
  - —small sample sizes
  - —normal populations

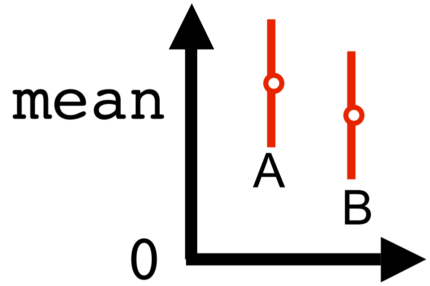
Approximate Visual Test

Simple visual test to compare unpaired samples

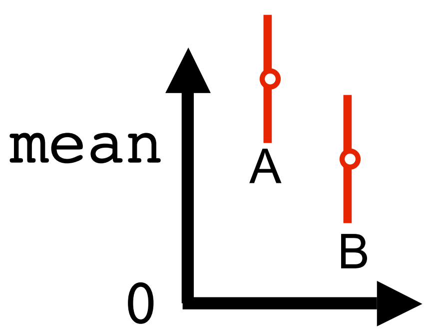
CI no overlap A > B



CI overlap; means in CI of other;
alternatives not different



CI overlap; mean A not in CI B;
need t-test



What Confidence Level to Use?

- Typically use confidence of 90% or 95%
- Need not always be that high
- Choice of confidence level is based on cost of loss if wrong!
- If loss is high compared to gain, use high confidence
- If loss is negligible compared to gain, low confidence OK

One-sided Confidence Intervals

- Sometimes only a one-sided confidence interval is needed
- Example: want to test if mean > μ_0
- In this case, one-sided lower confidence interval for μ needed

$$(\bar{x} - t_{[1-\alpha;n-1]} s/\sqrt{n}, \bar{x})$$

- For large samples, use z-values (unit normal distribution) rather than t-distribution

Confidence Intervals for Proportions

- For categorical variables, data often associated with probabilities of various categories
- Example: want confidence interval for n_1 of n of type 1
- Proportion = p = n_1/n
- Confidence interval for proportion

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Confidence interval based on approximating the binomial distribution
  —valid only if np ≥ 10
- If np ≤ 10, can't use t-values; must use binomial tables

Example: Confidence for Proportions

- **If 10 out of 1000 pages from a laser printer are illegible**
- **Proportion of illegible pages is 10/1000 = .01**
- **Since np ≥ 10, we can use the aforementioned confidence interval**

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$.01 \pm z_{1-\alpha/2}\sqrt{\frac{.01(.99)}{1000}}$$

Recall $z_{1-\alpha/2}$ is approximately $4.91[(1-\alpha/2)^{0.14} - (\alpha/2)^{0.14}]$

$$.01 \pm z_{1-\alpha/2}\sqrt{.0000099} = .01 \pm z_{1-\alpha/2}(.0031)$$

90% confidence = $.01 \pm (1.645)(.0031) = (.0049, .0151)$

95% confidence = $.01 \pm (1.96)(.0031) = (.0039, .0161)$

Determining Sample Size

- **Confidence level from a sample depends on sample size**
  —the larger the sample, the higher the confidence
- **Goal: determine smallest sample yielding desired accuracy**

Sample Size for Determining Mean

- **For a sample size n, the 100(1-α)% confidence interval of μ is**

$$\bar{x} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

- **For a desired accuracy of r%, the confidence interval must be**

$$\bar{x} \pm \bar{x} \frac{r}{100}$$

- **Thus,** $z_{1-\alpha/2} \frac{s}{\sqrt{n}} = \bar{x} \frac{r}{100}$ **and** $n = \left\lceil \left(\frac{100 z_{1-\alpha/2} s}{r\bar{x}} \right)^2 \right\rceil$
- **In a preliminary test, sample mean of response time is 20 seconds and std dev. = 5 seconds. How many repetitions are needed to estimate the mean response time within 2s at 95% confidence? Required accuracy r = 2 in 20 = 10%**

$$n = \left\lceil \left(\frac{100 z_{1-\alpha/2} s}{r\bar{x}} \right)^2 \right\rceil = \left\lceil \left(\frac{100(1.96)(5)}{(10)(20)} \right)^2 \right\rceil = \lceil 24.01 \rceil = 25$$

Sample Size for Determining Proportions

- **Recall confidence interval for proportion** $p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- **To get an accuracy of r,** $p \pm r = p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- **Thus,** $r = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ **and** $n = \left[\left(z_{1-\alpha/2}\right)^2 \frac{p(1-p)}{r^2}\right]$
- **A preliminary measurement of a laser printer showed an illegible print rate of 1 in 10000. How many pages must be observed to get an accuracy of 1 per million at 95% confidence?**

$$p = 1/10{,}000 = 10^{-4}; r = 10^{-6}; z_{.975} = 1.96$$

$$n = \left[(1.96)^2 \frac{10^{-4}(1-10^{-4})}{\left(10^{-6}\right)^2}\right] = 384{,}160{,}000$$