BIM 105
Probability and Statistics for Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

Point and Interval Estimation

- A population may sometimes be characterized by its *parameters*; for example, a normal population is completely described by the mean and the variance.
- At least, these parameters may be important, and it may be a goal to estimate them.
- A *statistic* is something calculated from data.
- We can sometimes view a statistic as an estimate of a parameter.
- For example, the mean μ of a population is a parameter. The sample mean $\overline{x}$ is such a statistic.

Estimators and Estimates

- An *estimator* is a method of computing a number from data, The result is called an *estimate*.
- If we are trying to estimate a parameter, it is best if the estimate is close to the parameter value.
- Since the estimate is a random variable, “close” has to be defined as an average distance from the true value, in some sense of average.
- This will depend on the estimator, the population, and the sample size, among other things.

Figures of Merit

Suppose we have a population with parameter θ

We have a random variable $X = \hat{\theta}$ that is meant to estimate θ

We would like to choose X from among available alternatives

so that X is on the average as close as possible to θ

1. We want X not to be too far to one side or the other

   $E(X-\theta) = \mu_X - \theta \approx 0$ Bias

2. We want the distance between X and θ not to be too large

   $E(X-\theta)^2$ is small

$$E(X-\theta)^2 = E(X-\mu_X+\mu_X-\theta)^2 = E(X-\mu_X)^2 - 2E(X-\mu_X)(\mu_X-\theta) + E(\mu_X-\theta)^2$$

$$= V(X) + E(\mu_X-\theta)^2$$

The mean square error of an estimate is the variance plus the square of the bias.

The Sample Mean

Suppose that $X_1, X_2, \ldots, X_n$ are a random sample from a population and that

$$E(X_i) = \mu_X$$

$$V(X_i) = \sigma_X^2$$

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$$

We previously saw that

$$E(\bar{X}) = \mu_X$$

$$V(\bar{X}) = \sigma^2 / n$$

So the bias of $\bar{X}$ is

$$E(\bar{X} - \mu_X) = \mu_X - \mu_X = 0$$

and the variance (and MSE) of $\bar{X}$ is

$$V(\bar{X}) = \sigma^2 / n$$

The Sample Proportion

Suppose that $X \sim \text{Bin}(n, p)$ and consider $\hat{p} = X / n$

We previously saw that

$$E(\hat{p}) = np / n = p$$

$$V(\hat{p}) = np(1 - p) / n^2 = p(1 - p) / n$$

So the bias of $\hat{p}$ is

$$E(\hat{p} - p) = p - p = 0$$

and the variance (and MSE) of $\hat{p}$ is

$$V(\hat{p}) = p(1 - p) / n$$

Unbiased vs. Minimum MSE

- Unbiased sounds like a good thing, but it is not the most important characteristic.
- A small amount of bias relative to the variability of the estimator may not be important.
- The sample variance s^2 (with denominator $n - 1$) is an unbiased estimate of the population variance σ^2.
- But the sample standard deviation s is not unbiased for the population standard deviation σ.
- We use it anyway, because it is optimal in other ways.

Consistency/Convergence

Another important property of a point estimator is that

as the sample size gets larger, the estimate gets closer to the true value.

This is called *consistency*.

For our purposes, this simply means that the MSE $\to 0$ as $n \to \infty$

For the sample mean

$$\text{MSE}(\bar{X}) = \sigma^2 / n \to 0$$

For the sample proportion

$$\text{MSE}(\hat{p}) = p(1-p)/n \to 0$$

Interval Estimates

The idea of an interval estimate is to provide an interval $[a,b]$ that is "likely" to contain the unknown parameter value θ. Suppose we are interested in the calcium content of a cell-growth medium, and in particular, we want to know that the mean content is across batches of the medium. Suppose we take a sample of size 100 and the sample mean calcium content is 36 gm/L with a sample standard deviation of 14 gm/L.

We know the sample mean has standard deviation (also call standard error of the mean)

$\sigma / \sqrt{100} \doteq 14/10 = 1.4.$

What values of the population mean μ are consistent with this evidence?

$\mu = 38$ or $\mu = 34$ are consistent with the findings.

$\mu = 42$ or $\mu = 30$ are not consistent with the findings.

Confidence Intervals

- A 95% confidence interval is an interval such that the probability that the true parameter value is in the interval is 0.95.
- We can also define 99% confidence intervals or, in general, the 100(1 – α)% confidence interval for any value of $0 < \alpha < 1$.
- The usual interpretation of this statement is based on the idea that the parameter has a fixed value that we do not know.
- We take a sample (which is random) and calculate the interval (which is random) and if we repeated this procedure, some of the intervals contain the true parameter value and some don't.

- In such repeated sampling, 95% of the resulting intervals are supposed to contain the true parameter value.
- This is called the *frequentist* approach since it is based on hypothetical repeats of the experiment and a statement about how frequently the interval will contain the true parameter value.
- An alternative approach is to treat the parameter value as itself a random variable (which we do not observe). In this *Bayesian* approach, the interval is based on the data, and then the statement that the parameter value is in the interval is a probability statement about the random parameter value.
- Much of the time, the two approaches yield very similar results.
- We will concentrate on the first approach, which is the more usual in science and engineering.

The calcium content across batches of a cell-growth medium has mean μ and variance σ^2. The sample mean $\overline{x}$ of a sample of size n has

$$E(\overline{x}) = \mu$$

$$V(\overline{x}) = \sigma^2 / n$$

From the central limit theorem, $\overline{x}$ is approximately normally distributed so that

$$\overline{x} \sim N(\mu, \sigma^2 / n)$$

$$\frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$P(-z_{.025} < \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} < z_{.025}) = 0.95$$

$$P(\mu - z_{0.025}\sigma / \sqrt{n} < \overline{x} < \mu + z_{0.025}\sigma / \sqrt{n}) = 0.95$$

$$P(\mu - z_{0.025}\sigma / \sqrt{n} < \overline{x} < \mu + z_{0.025}\sigma / \sqrt{n}) = 0.95$$

The event above is that both the following are true:

$\overline{x} < \mu + z_{0.025}\sigma / \sqrt{n}$ and $\mu - z_{0.025}\sigma / \sqrt{n} < \overline{x}$ but

$$P(\overline{x} < \mu + z_{0.025}\sigma / \sqrt{n}) = P(-\mu < -\overline{x} + z_{0.025}\sigma / \sqrt{n}) = P(\mu > \overline{x} - z_{0.025}\sigma / \sqrt{n})$$

$$P(\mu - z_{0.025}\sigma / \sqrt{n} < \overline{x}) = P(-\overline{x} - z_{0.025}\sigma / \sqrt{n} < -\mu) = P(\overline{x} + z_{0.025}\sigma / \sqrt{n} > \mu)$$

$$0.95 = P(\mu - z_{0.025}\sigma / \sqrt{n} < \overline{x} < \mu + z_{0.025}\sigma / \sqrt{n}) = P(\overline{x} - z_{0.025}\sigma / \sqrt{n} < \mu < \overline{x} + z_{0.025}\sigma / \sqrt{n})$$

$$P(\overline{x} - z_{0.025}\sigma / \sqrt{n} < \mu < \overline{x} + z_{0.025}\sigma / \sqrt{n}) \approx P(\overline{x} - z_{0.025}s / \sqrt{n} < \mu < \overline{x} + z_{0.025}s / \sqrt{n})$$

Suppose we take a sample of size 100 and the sample mean calcium content is 36 gm/L with a sample standard deviation of 14 gm/L. Then a 95% confidence interval is

$$36 \pm (1.960)(14) / \sqrt{100} = 36 \pm (1.960)(1.40) = 36 \pm 2.744$$

$$(33.256, 38.744)$$

Large-Sample Intervals for the Mean

A large-sample $100(1-\alpha)\%$ confidence interval for the mean μ

based on the sample mean $\overline{x}$ and sample standard deviation s of a sample of size n is

$$\overline{x} \pm z_{\alpha/2} s/\sqrt{n}$$

The statistic $s/\sqrt{n}$ is called the *standard error* of $\overline{x}$.

As against the statistic s which is the *sample standard deviation.*

If we repeat the process of taking a sample of size n and computing the confidence interval, then the fraction of the time that the interval contains μ is $(1-\alpha)$.

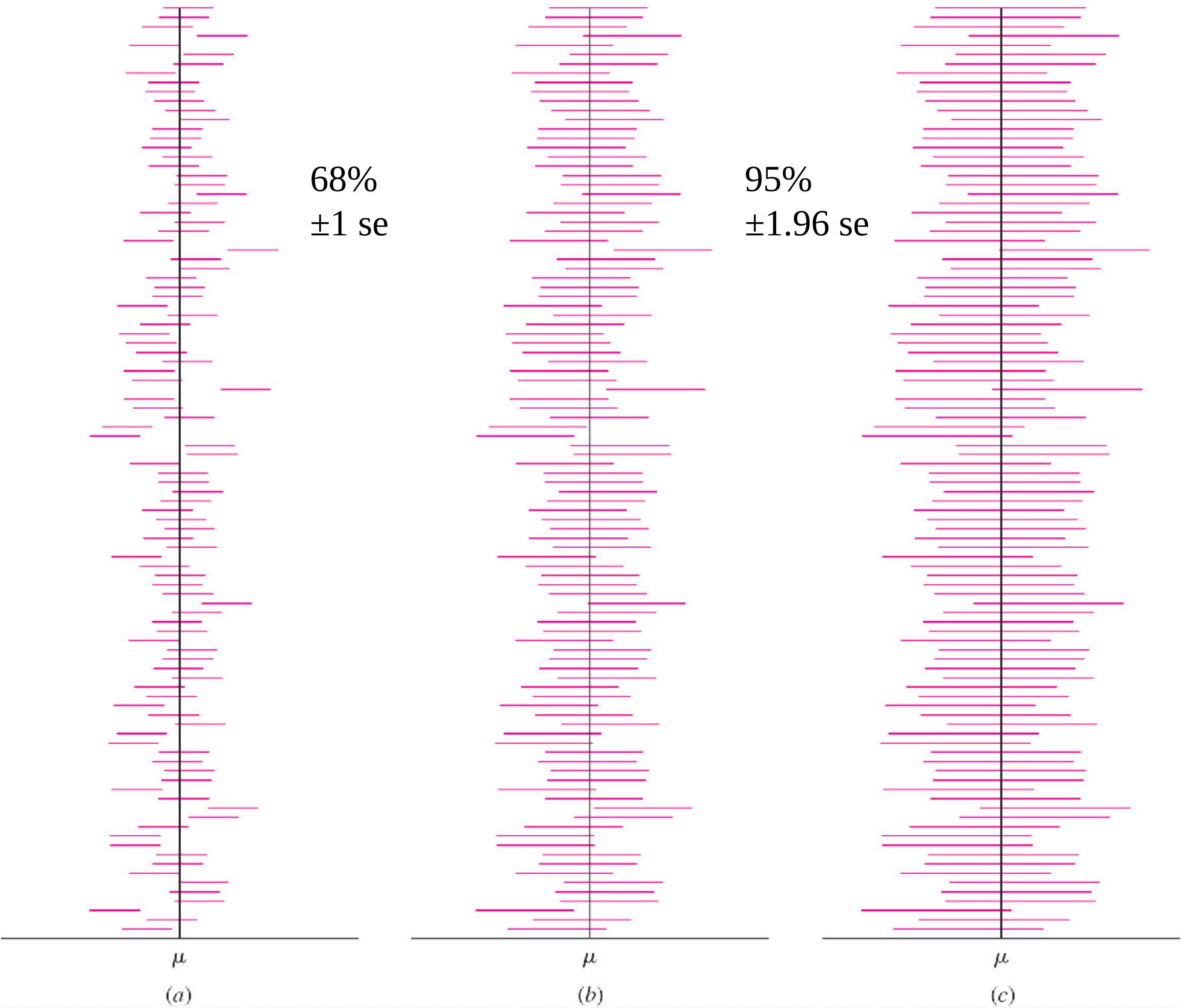
CI	50%	90%	95%	99%	99.9%
z_α	0.6745	1.645	1.960	2.576	3.291

Confidence Interval for the IgM Data

- There are 298 observations in the IgM data set (concentrations of IgM in g/L).
- Since the distribution looks more normal on the log scale, we do the analysis on that scale.
- This gives us an interval for the mean of ln(IgM), which is the geometric mean of IgM on the original scale.
- The mean is –0.3632 and the standard deviation is 0.5469
- A 95% CI is
  $-0.3632 \pm (1.960)(0.5469)/\sqrt{298}$
  $= -0.3632 \pm (1.960)(0.0317)$
  $= -0.3632 \pm 0.0621$
  $(-0.4253, -0.3011)$
- Or (0.653, 0.740) on the original scale of g/L, which we obtain by exponentiating each end of the CI.

Behavior of Confidence Intervals

- A 95% confidence interval will cover the true value 95% of the times that the procedure is run.
- A particular 95% confidence interval either covers the true value or not, though we don’t know which in an particular case.
- So (from the frequentist point of view) it is incorrect to say that the probability that the interval covers the true value is 95%.
- That is why we call it 95% confidence and not 95% probability.



68%
±1 se

95%
±1.96 se

99.7%
±3 se

Small-Sample Confidence Intervals for the Mean

The usual large-sample interval for the mean is based on the fact that

$$E(\bar{X}) = \mu$$

$$V(\bar{X}) = \sigma^2 / n$$

$\bar{X}$ is approximately normally distributed, so that

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

which is equivalent to saying that

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

If n is large, then replacing the population variance σ, which we don't know by the sample standard deviation s does not change this much.

If n is small, then replacing a known constant denominator by a random variable will make the ratio more variable, so that the standard deviation is no longer 1.

It turns out that if X is normally distributed, then

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

The t distribution, or Student's t distribution has mean 0,

but is more variable than the standard normal

$$V(t_\nu) = \frac{\nu}{\nu - 2}$$

To construct a confidence interval for the mean using the t distribution,

we replace the normal percentage point with a t percentage point.

$$\bar{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$$

We can get these percentage points from Table A.3 on page 523

Or we can use the MATLAB command `tinv()`

For example, for a 95% confidence interval with n = 10, we have $\nu = 9$

Table A.3 has the 0.025 upper percentage point of a t with 9df as 2.262

```
>> tinv(.975,9)

     2.2622
```

Behavior of the t-Distribution

- The t-distribution gets closer to the normal as n gets larger.
- If n is large, one can use the normal percentage point instead of the t percentage point.
- But using computer analysis, there is never any harm in using the t, so that is the default in MATLAB for confidence intervals for the mean.

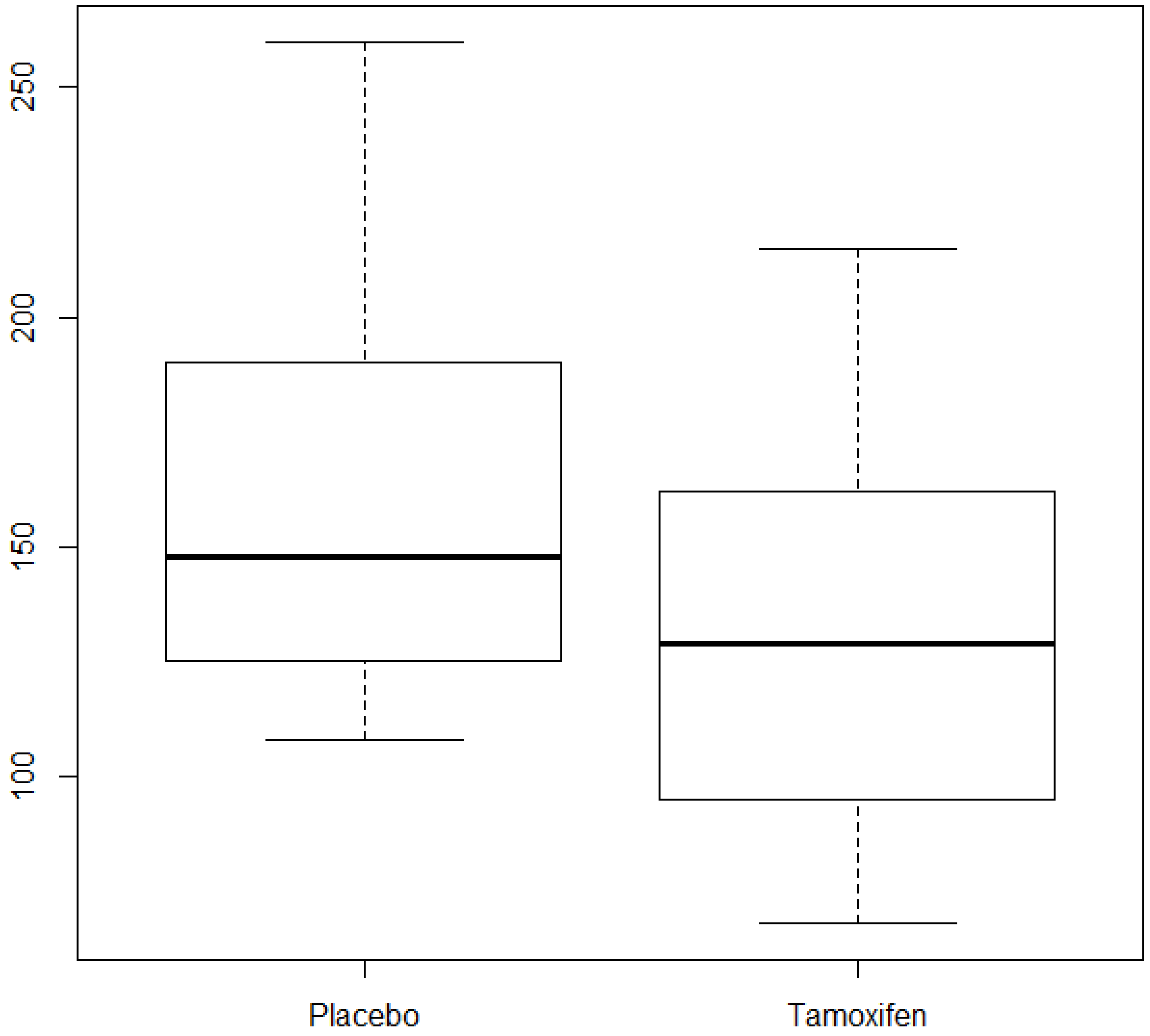
n	5	10	30	100	200	1000	5000
ν	4	9	29	99	199	999	4999
t(0.025)	2.776	2.262	2.045	1.984	1.972	1.962	1.960

Does X have to be normal?

- The t-distribution is derived mathematically from the assumption that the population is normally distributed.
- Modest departures from this do not matter much.
- If the distribution is skew, then it may make sense to take logs before analysis.
- If there are large outliers, then these should be examined.
- There are robust and resistant versions of the t-distribution if the distribution is known to be outlier prone.

Alkaline phosphatase data

- Repeated measurements of alkaline phosphatase in a randomized trial of Tamoxifen treatment of breast cancer patients.
- We use the measurements from 24 months and make confidence intervals for the placebo (21 subjects) and Tamoxifen (17 subjects) groups.
- Elevated alkaline phosphatase can be a sign of recurrence or metastasis.


250
200
150
100
Placebo
Tamoxifen

For the placebo group

$n = 21$

$\nu = 20$

$\bar{x} = 163.29$

$s = 46.03$

$t_{20,0.025} = 2.0860$

$\bar{x} \pm t_{20,0.025} s / \sqrt{21}$

$163.29 \pm (2.0860)(46.03) / \sqrt{21}$

163.29 ± 20.95

$(142.34, 184.24)$

For the treatment group

$n = 17$

$\nu = 16$

$\bar{x} = 133.71$

$s = 47.21$

$t_{16,0.025} = 2.1199$

$133.71 \pm (2.1199)(47.21) / \sqrt{17}$

133.71 ± 24.27

$(109.44, 157.98)$

Confidence Intervals in MATLAB

```
>> igmdist = fitdist(ligm,'Normal')

NormalDistribution

  Normal distribution
       mu = -0.363163   [-0.425511, -0.300816]
    sigma =  0.546895   [0.506229, 0.594721]
>> paramci(igmdist)

   -0.4255    0.5062
   -0.3008    0.5947
```

Confidence Intervals in MATLAB

```
>> [h p ci stats] = ttest(ligm)  % ignore h and p. This is a test that the
                                 % mean of the log IgM is 0, which is not
h =                              % meaningful

     1

p =

   1.9612e-25

ci =

   -0.4255
   -0.3008

stats =

    tstat: -11.4632
       df: 297
       sd: 0.5469
```

Confidence Intervals in MATLAB

```
Confidence interval for mean log IgM

ci =
   -0.4255
   -0.3008

Confidence interval for geometric mean IgM

>> exp(ci)

ans =

    0.6534
    0.7402
```

Sample Size Determination

- In the calcium content example, we have a sample of size 100 with mean 36 and standard deviation 14.
- We got a 95% confidence interval of 36 ± 2.744.
- Suppose we needed to know the mean concentration to within ±1 gm/L (that is, that the 95% CI would have the form 36 ± 1). How big must *n* be?
- $1.0 = (1.960)(14)/\sqrt{n}$

  $n = (1.960)^2(14)^2/1.0^2$

  $n = 752.95$

  $n = 753$

For the $100(1-\alpha)\%$ CI to have width w

$\overline{x} \pm w$

The sample size n must be at least

$$n = \frac{z_{\alpha/2}\sigma^2}{w^2}$$

This should be rounded up to the nearest integer.

An Interpretation of CI’s

Another perspective on confidence intervals is to ask the following question: given a sample mean $\overline{x}$, sample variance s^2, and a possible population mean μ, is the possible value μ for the mean consistent with the data?

If μ is too far from $\overline{x}$, then we would say no.

The distance from $\overline{x}$ to μ is $|\overline{x}-\mu|$

If that distance is bigger than $1.96s/\sqrt{n}$, then we would perhaps say no.

$\dfrac{\overline{x}-\mu}{s/\sqrt{n}} > 1.96$ means that μ is not consistent with the data.

If true mean is μ, then $\overline{x} \sim N(\mu, \sigma/\sqrt{n})$

$\Pr(|\overline{x}-\mu| > 1.960\sigma/\sqrt{n}) = .05$

The 95% CI is the set of possible values of μ that are consistent with the data with a 5% criterion for not consistent.

Confidence Intervals for Proportions

- Large sample intervals for proportions are also based on the standard error of the statistic (sample proportion) and on normal percentage points (because of the central limit theorem.
- The variance of a sample proportion is $p(1 - p)/n$
- So one possible approach to a 95% confidence interval is to substitute the sample proportion in this formula for the parameter p.
- This could be called the traditional method, but there are other possibilities.

Estimating the Binomial Proportion

If we get x successes out of n, then it makes sense to estimate p as

$$\hat{p} = \frac{x}{n}$$

But this is not the only possibility.

If we try inducing stem-cell behavior in 10 fibroblast cells, is $p = 0$ the best estimate?

If we test 10 medical devices and all 10 function correctly, is $p = 1$ the best estimate?

Other commonly used estimates can be justified by Bayesian reasoning:

$$\hat{p} = \frac{x + 1/2}{n + 1}$$

$$\hat{p} = \frac{x + 1}{n + 2}$$

$$\hat{p} = \frac{x + 2}{n + 4}$$ this one is described in the book

These all behave well at the ends. With 0/10 we get 1/22, 1/12, or 1/7.

With 10/10 we get 21/22, 11/12, or 6/7.

When we developed the 95% CI for the mean, we used the fact that the statement that $\overline{x}$ has a 95% chance of lying in the interval

$$\mu \pm 1.960\sigma / \sqrt{n}$$

which can be inverted to state that the interval

$$\overline{x} \pm 1.960\sigma / \sqrt{n}$$

has a 95% chance to contain μ. Here σ is separately estimated and does not depend on μ.

For the binomial, the statement that $\hat{p}$ has a 95% chance to lie in

$$p \pm 1.960\sqrt{p(1-p)/n}$$

is only approximately true because of discreteness and the fact that the CLT is approximate. If we try to invert this statement, to obtain that p lies 95% of the time in

$$\hat{p} \pm 1.960\sqrt{\hat{p}(1-\hat{p})/n}$$

it does not quite work, because $\mathrm{Var}(\hat{p})$ changes with p.

The traditional 95% CI for p is

$$\hat{p} \pm 1.960\sqrt{\hat{p}(1-\hat{p})/n}$$

using

$$\hat{p} = \frac{x}{n}$$

The book advocates for a different version that

uses a different center and variance estimate.

$$\tilde{p} \pm 1.960\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

$$\tilde{n} = n + 4$$

$$\tilde{p} = \frac{x+2}{\tilde{n}}$$

For exercises from the book, use this; it has some advantages.

The traditional interval is more common in the wild.

Both procedures are examples of Wald intervals,

a general procedure based on a normal approximation.

Another approach, and the default in MATLAB, is the Clopper-Pearson interval or so-called exact interval.

If we have x successes out of n then

$\hat{p} = x/n$

$V(\hat{p}) \approx \hat{p}(1-\hat{p})/n$

When is a possible value of p_0 not consistent with the data?

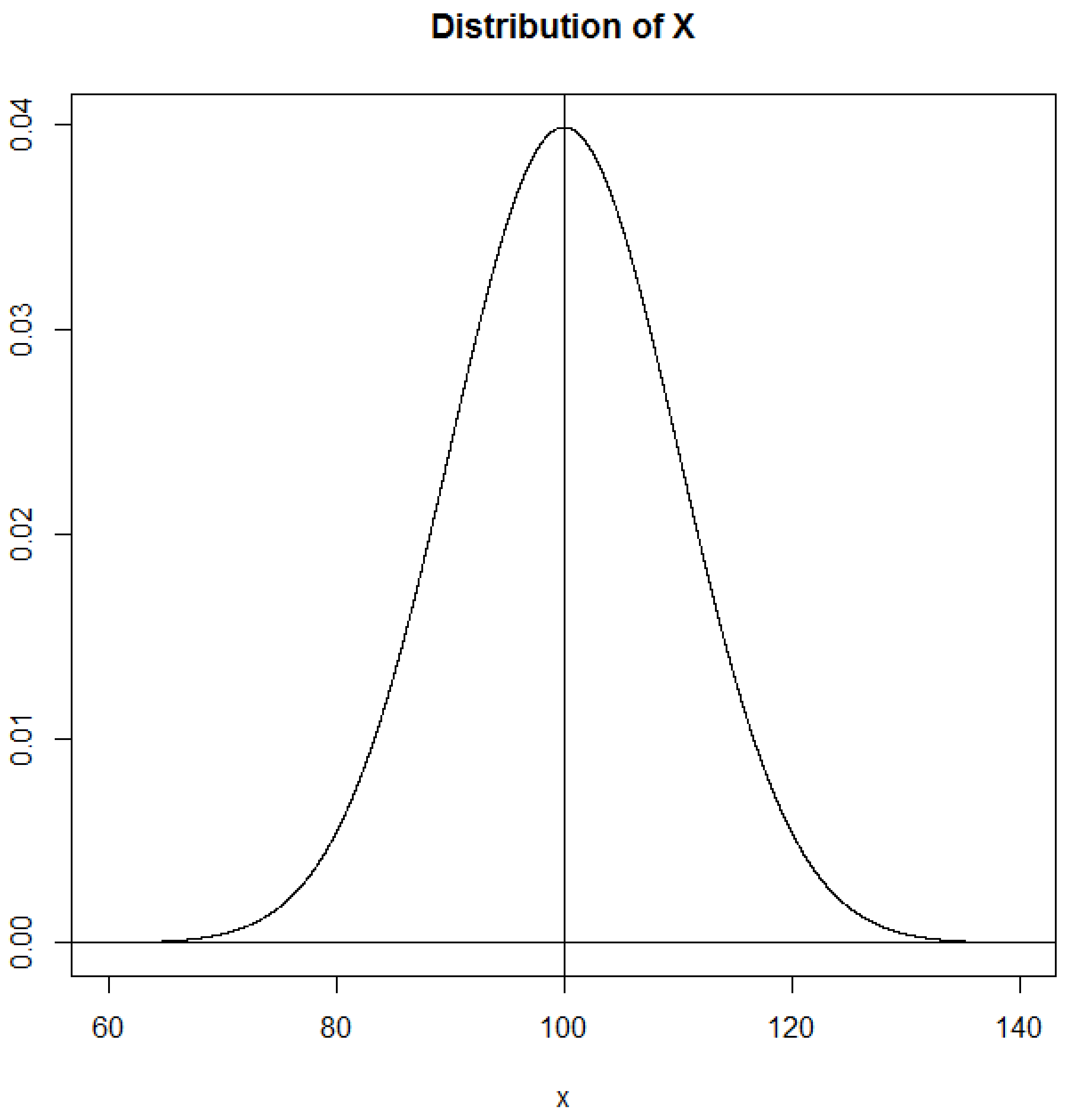
For the moment, suppose that the hypothesized value p_0 is larger than $\hat{p}$

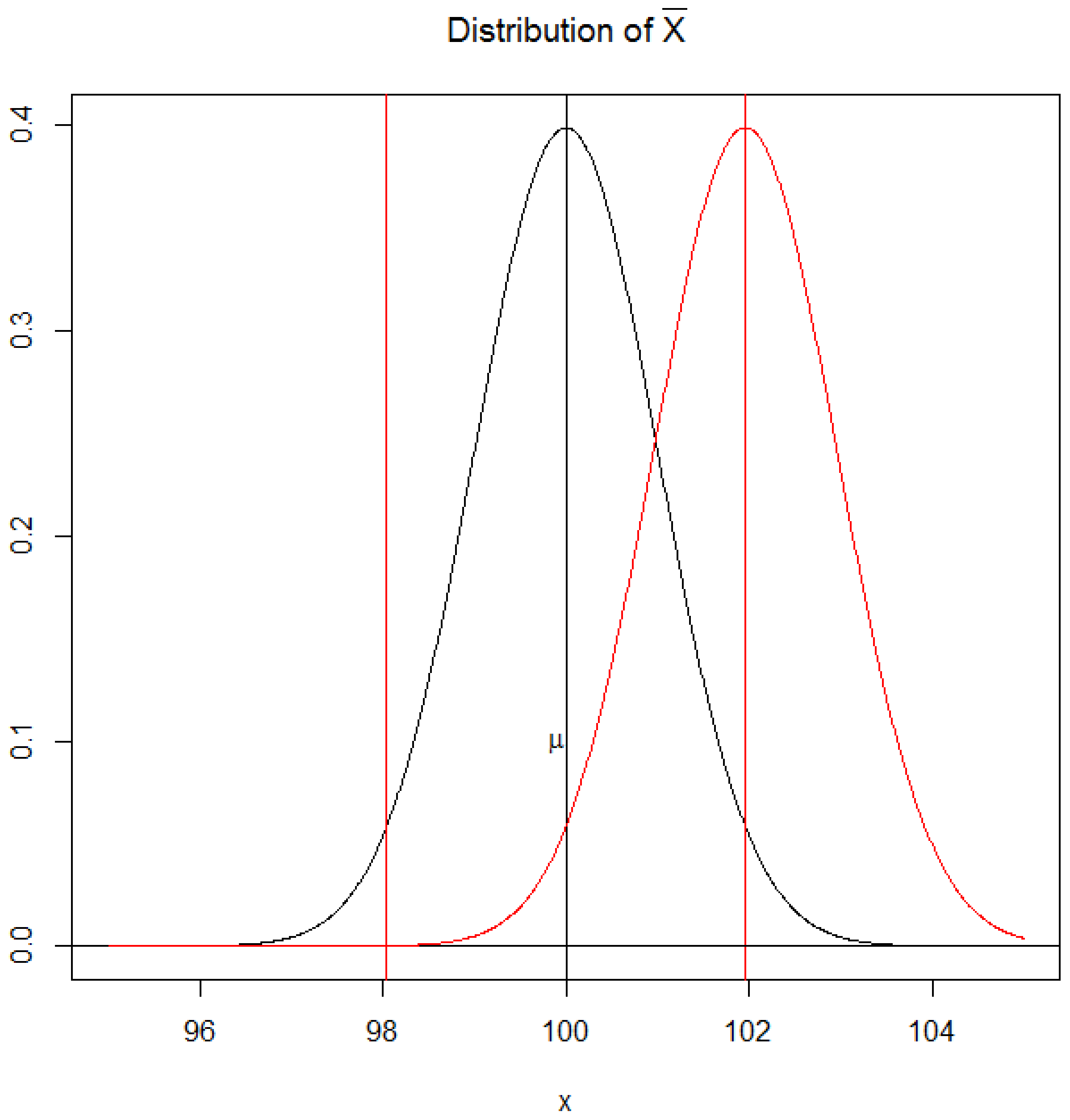
Compute $\Pr(X \le x \mid p = p_0)$. If this is smaller than 0.025, then say that that value of p is not consistent. Note that when we are checking whether p_0 is consistent, we have

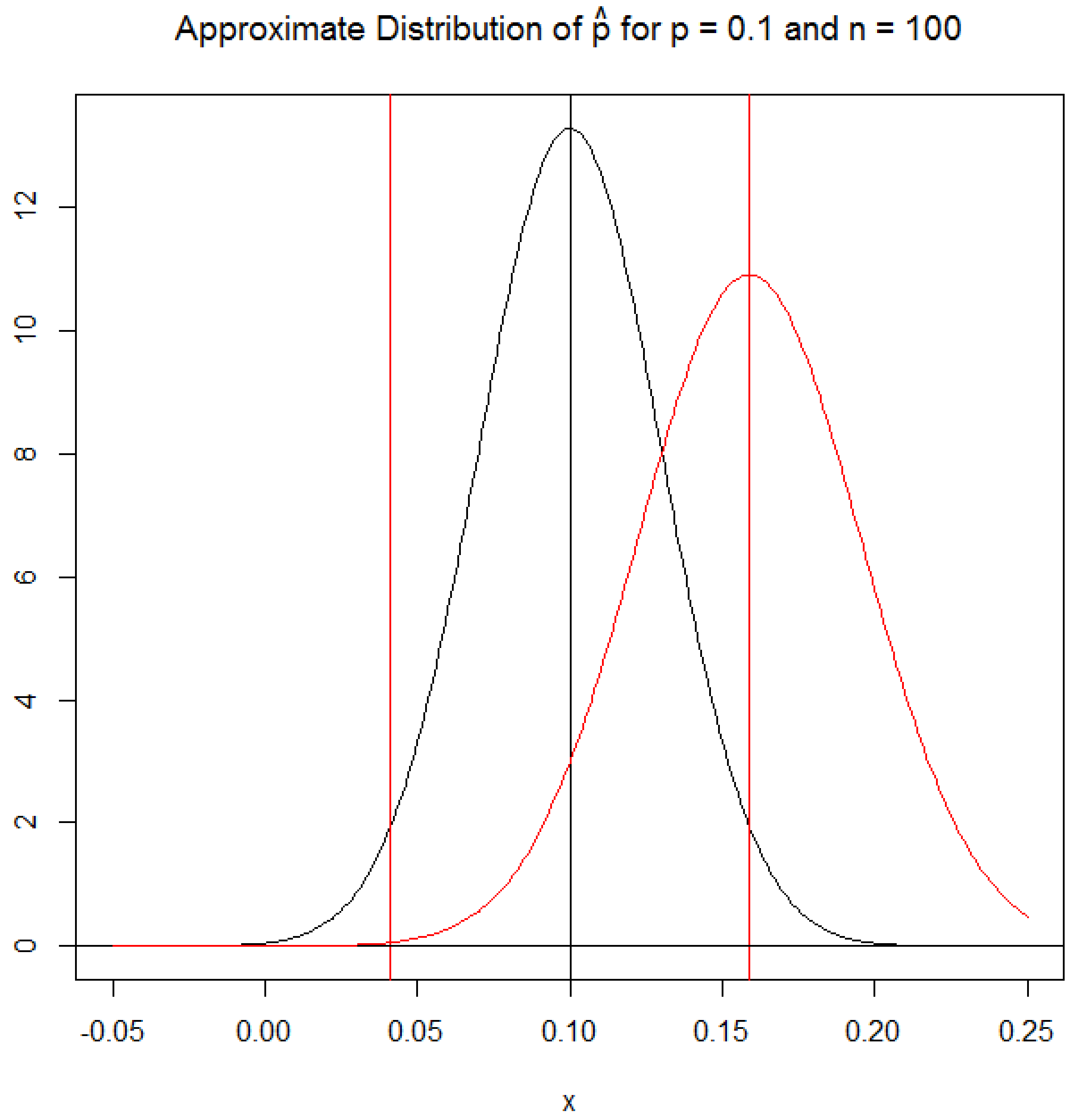
$\mathrm{V}(\hat{p}) = p_0(1-p_0)/n \ne \hat{p}(1-\hat{p})/n$

The value of p_0 that makes $\Pr(X \le x \mid p = p_0) = 0.025$ forms the upper end of the 95% CI in this method, and similarly for the lower.

We will check this out more later in the lecture when we look at MATLAB


Distribution of X
0.04
0.03
0.02
0.01
0.00
60
80
100
120
140
x


Distribution of $\overline{X}$
0.0
0.1
0.2
0.3
0.4
96
98
100
102
104
x
μ


Approximate Distribution of $\hat{p}$ for p = 0.1 and n = 100
0
2
4
6
8
10
12
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
x

We want to evaluate a new method of inducing iPSC stem cells.

In a trial of *n* = 100 cells, 14 were successfully transformed.

Find a 95% confidence interval for the true proportion transformed.

$X = 14$

$n = 100$

The usual (Wald) method using the normal approximation

$\hat{p} = 14/100 = 0.14$

$\text{sd}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = 0.0347$

$0.14 \pm (1.960)(0.0347) = (0.0720, 0.2081)$

The Agresti-Coull method that is presented in the book gives

$\tilde{p} = 16/104 = 0.1539$

$\text{sd}(\tilde{p}) = \sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}} = 0.0354$

$0.1539 \pm (1.960)(0.0354) = (0.0846, 0.2232)$

Sample Size Determination

If we have a desired CI half width of $\pm w$

and if we have a prior estimate of p

then n needs to be at least large enough so that

$$w = z_{\alpha/2}\sqrt{p(1-p)/n}$$

$$w^2 = z_{\alpha/2}^2 p(1-p)/n$$

$$n = z_{\alpha/2}^2 p(1-p)/w^2$$

For example if $p = 0.14,$ and we want a half-width of 0.05

for 95% confidence, then the required sample size is

$$n = (1.960)^2(0.14)(0.86)/(0.05)^2 = 185.01$$

so we need a sample size of 186

Example

- The birth weight of 189 infants was collected at Baystate Medical Center in Springfield, MA.
- The mean weight was 2945 gm with a standard deviation of 729 gm.
- A 95% confidence interval for the mean birth weight is $2945 \pm (1.960)(729)/\sqrt{189} = 2945 \pm 104 = (2841, 3049)$
- 59 out of the 189 infants had birth weights below the safe level of 2.5 kg, which is 31.2%.

- 59 out of the 189 infants had birth weights below the safe level of 2.5 kg.
  - p = 0.3122.
  - SD(p) = √(0.3122)(0.6888)/189 = √ 0.00114 = 0.0337
- The usual 95% CI is
  - 0.3122 ± (1.960)(0.0337)
  - 0.3122 ± 0.0661
  - (0.246, 0.378)

Cell Migration Study

- Keratinocytes were plated at a density of 50 cells/mm^2 onto collagen-coated coverslips for 2 h at 37 °C.
- To simulate the high epinephrine environment seen in burn patients , culture medium was supplemented with epinephrine.
- There was also a control in which the epinephrine was not added.
- The distance moved by the cells in one hour was measured.

- For the control, the mean distance of 24 cells was 1.189 mm with a standard deviation of 0.883 mm
- A 90% confidence interval for the mean is then
  $1.189 \pm (1.645)(0.883/\sqrt{24})$
  $1.189 \pm (1.645)(0.180)$
  1.189 ± 0.296
  $(0.893, 1.485)$
- For the treated cells, the mean distance of 22 cells was 0.772 mm with a standard deviation of 0.758 mm.
- A 90% confidence interval for the mean is then
  $0.772 \pm (1.645)(0.758/\sqrt{22})$
  $0.772 \pm (1.645)(0.162)$
  0.772 ± 0.266
  $(0.506, 1.038)$

Confidence Intervals in MATLAB Tamoxifen data using t interval

```
>> placebo'
   148   116   221   189   114   218   190   123   138   232
260   111   228   172   147   150   146   108   129   125   164

>> placebo_summary = fitdist(placebo,'Normal')

  Normal distribution
       mu = 163.286   [142.334, 184.238]  #this is the t interval
    sigma = 46.0284   [35.2145, 66.4682]

>> paramci(placebo_summary)

  142.3338   35.2145    # first column is t interval for mean
  184.2376   66.4682    # second column is interval for sigma
```

Confidence Intervals in MATLAB Stem cell data, intervals for p

```
>> stemdist = fitdist(14,'Binomial','NTrials',100)

  BinomialDistribution

  Binomial distribution
    N =  100
    p = 0.14   [0.0787054, 0.223728]
>> paramci(stemdist)

  100.0000    0.0787
  100.0000    0.2237
>> paramci(stemdist,'Type','Wald')

  100.0000    0.0720
  100.0000    0.2080
```

This an ‘exact’ interval whose coverage is conservative (> 95%)

This is the traditional interval whose coverage is liberal (< 95%)

```
>> binocdf(14,100,.20)
    0.0804
>> binocdf(14,100,.22)
    0.0305
>> binocdf(14,100,.23)
    0.0177
>> binocdf(14,100,.225)
    0.0233
>> binocdf(14,100,.223)
    0.0260
>> binocdf(14,100,.224)
    0.0246
>> binocdf(14,100,.2235)
    0.0253
>> binocdf(14,100,.2237)
    0.0250
>> paramci(stemdist)
  100.0000    0.0787
  100.0000    0.2237
```

Summary

For a large-sample $100(1-\alpha)\%$ confidence interval for the mean of measured data, we use

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

For binomial data, if we have x successes out of n in a sample then a $100(1-\alpha)\%$ confidence interval for the true proportion of successes p is

$$\hat{p} \pm \sqrt{\hat{p}(1-\hat{p})/n}$$

where usually $\hat{p} = x/n.$

MATLAB calls this procedure the Wald interval. The default in MATLAB is the Clopper-Pearson so-called exact interval. It is exact in the sense that it uses the binomial cdf instead of the normal approximation. It is not exact in the sense that the coverage is actually exactly 95%. In practical use, either method is acceptable, as is the alternate method in the book, the Agresti-Coull interval.

Summary

To obtain a confidence interval with half-width *w*

$\overline{x} \pm w$ or

$\hat{p} \pm w$

the required sample size *n* is

$$n = \frac{z_{\alpha/2}\sigma^2}{w^2}$$

for the mean of measured data and

$$n = z_{\alpha/2}^2 p(1-p) / w^2$$

for the proportion

both rounded up to the nearest whole number