BIM 105
Probability and Statistics for Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

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.

Confidence Intervals in MATLAB

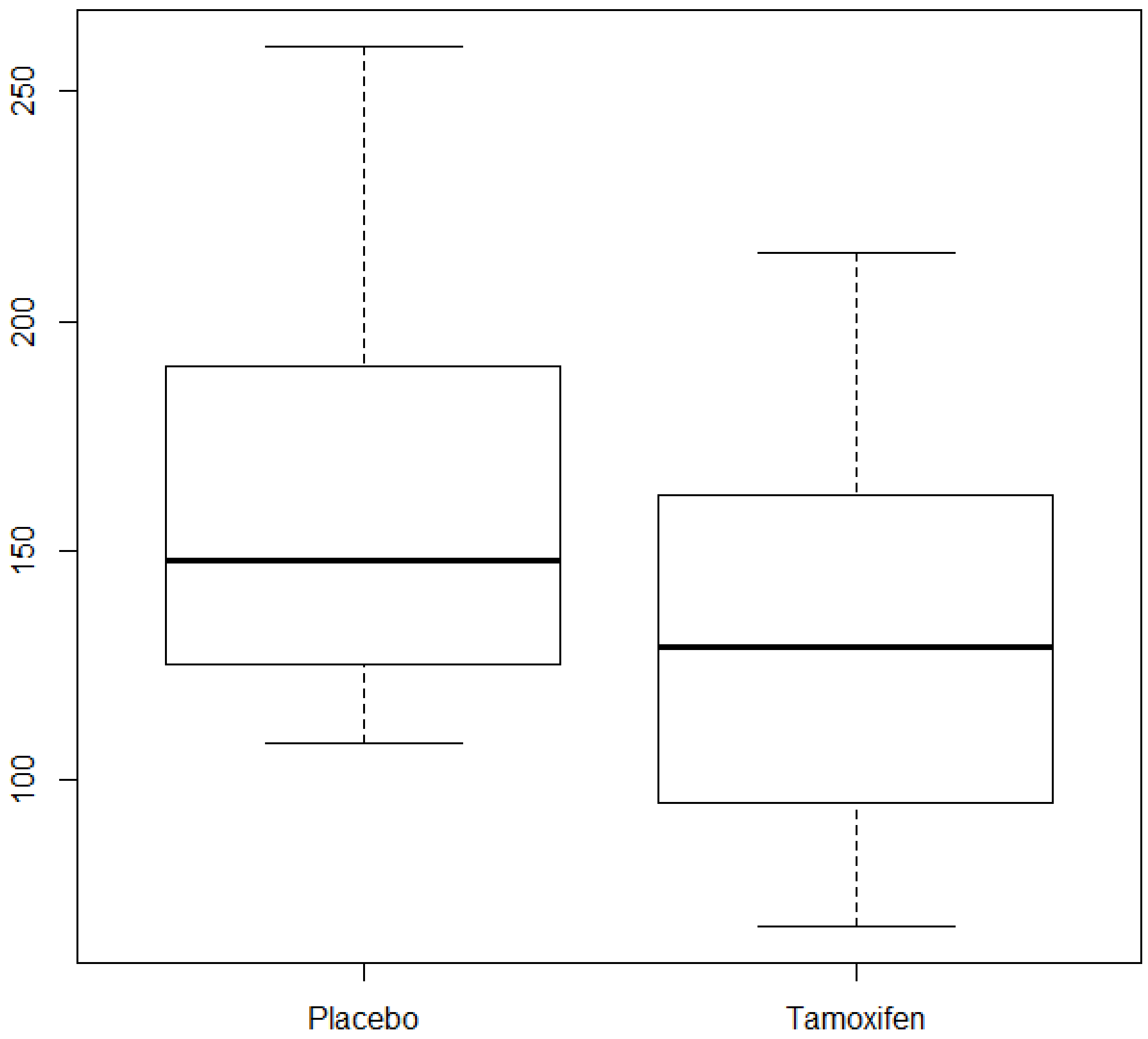
```
>> [h p ci stats] = ttest(ligm)  % ignore h and p. This is a test that the
                                 % mean of the log IgM is 0, which is not
h =                              % meaningful

     1

p =

   1.9612e-25

ci =

   -0.4255            %Compare to z-interval in last lecture
   -0.3008            %differs only in the fourth significant figure

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
```

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

$$n = 21$$

$$\nu = 20$$

$$\overline{x} = 163.29$$

$$s = 46.03$$

$$t_{20,0.025} = 2.0860$$

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

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

$$163.29 \pm 20.95$$

$$(142.34, 184.24)$$

For the treatment group

$$n = 17$$

$$\nu = 16$$

$$\overline{x} = 133.71$$

$$s = 47.21$$

$$t_{16,0.025} = 2.1199$$

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

$$133.71 \pm 24.27$$

$$(109.44, 157.98)$$

Confidence Intervals in MATLAB Tamoxifen data using t interval

```
>> [h p ci t] = ttest(placebo)

h =
     1                  #ignore h and p---related to hypothesis
p =                      test
   5.4100e-13
ci =
  142.3338
  184.2376
t =
    tstat: 16.2567
       df: 20
       sd: 46.0284
```

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 − α)% CI to have width w

$\overline{x} \pm w$

The sample size n must be at least

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

This should be rounded up to the nearest integer.

One-Sample Tests

If we have a sample $x_1, x_2, ..., x_n$ from a population with mean μ

and we have a possible value $\mu = \mu_0$ in mind

and want to see if the data are consistent with that value of μ,

we construct the test statistic

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

and we compare it to the t-distribution with $n-1$ degrees of freedom,

or with the normal distribution if n is large.

If t is positive and $\Pr(T > t) = \alpha$,

then the p-value of the test is 2α.

If t is negative and $\Pr(T < t) = \alpha$,

then the p-value of the test is 2α.

For example, if 18 cells were sized

and the average diameter was 22 μm with a

standard deviation of 3 μm,

test the hypothesis that the true value of μ is 20 μm.

$$\frac{22-20}{3/\sqrt{18}} = \frac{2}{0.707} = 2.828 = t_{17}$$

$\Pr(T_{17} > 2.828) = 0.0058$ so the p-value is 2(0.0058)=0.0116.

The value $\mu = 20$ is not very consistent with the data.

This requires MATLAB to compute the tail area 0.0058.

From Table A.3, we can say that the tail area is less than 0.01 (2.567)

but not as small as 0.005 (2.898), so we can say only that p < 0.02.

Inference for Proportions

- Large sample test and 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 $p(1 - p)/n$
- If we are testing the null hypothesis H_o: $p = p_o$, then under the null we know that the sample proportion has mean p_o, variance $p_o(1 - p_o)/n$ and is approximately normally distributed.

Proportions

Suppose we have a sample from a production run of a component for a medical device and they are tested to see if they are within specifications. The manufacturing process is delicate, so up to 30% of the components may be defective. If we have 600 components of which 217 are defective, test the hypothesis that $p = 0.30$.

Note that

$$\hat{p} = X / n = 217 / 600 = 0.36167$$

Is the hypothesized value too far from the value estimated from the data?

Proportions

$$H_0 : p = 0.30$$
$$\hat{p} = X / n = 217 / 600 = 0.36167$$

For a large value of n, we have that

$X \sim \mathrm{Bin}(n, p) = \mathrm{Bin}(600, 0.30)$ if $p \;=\; p_0 = 0.30$

$E(\hat{p}) = p = 0.30$ under the null hypothesis

$$V(\hat{p}) = p(1-p)/n = (0.30)(0.70)/600 = 0.00035$$

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} = \frac{217/600-0.30}{\sqrt{(0.30)(0.70)/600}} = \frac{0.06167}{0.01871} = 3.296 = z$$

The p-value for this test is 2(0.00049) = 0.00098, so not consistent. We reject the null hypothesis.

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,

and none are converted, 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:

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

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

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

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

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

which can be inverted to state that the interval

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

has a 95% chance to contain μ. Here s 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 (consequently different) 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?

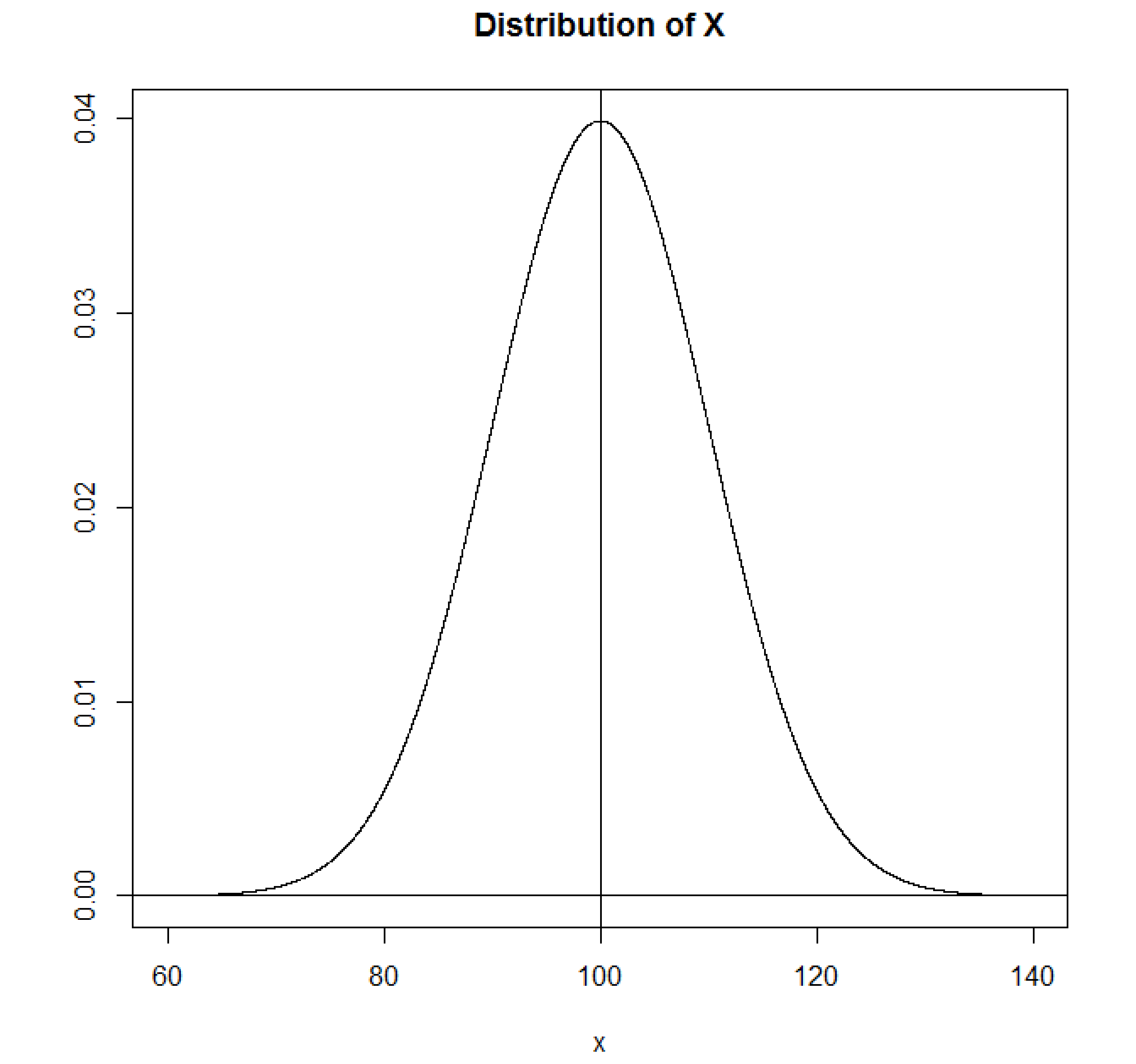
For the moment, suppose that the hypothesized value p_0 is larger than $\hat{p}$

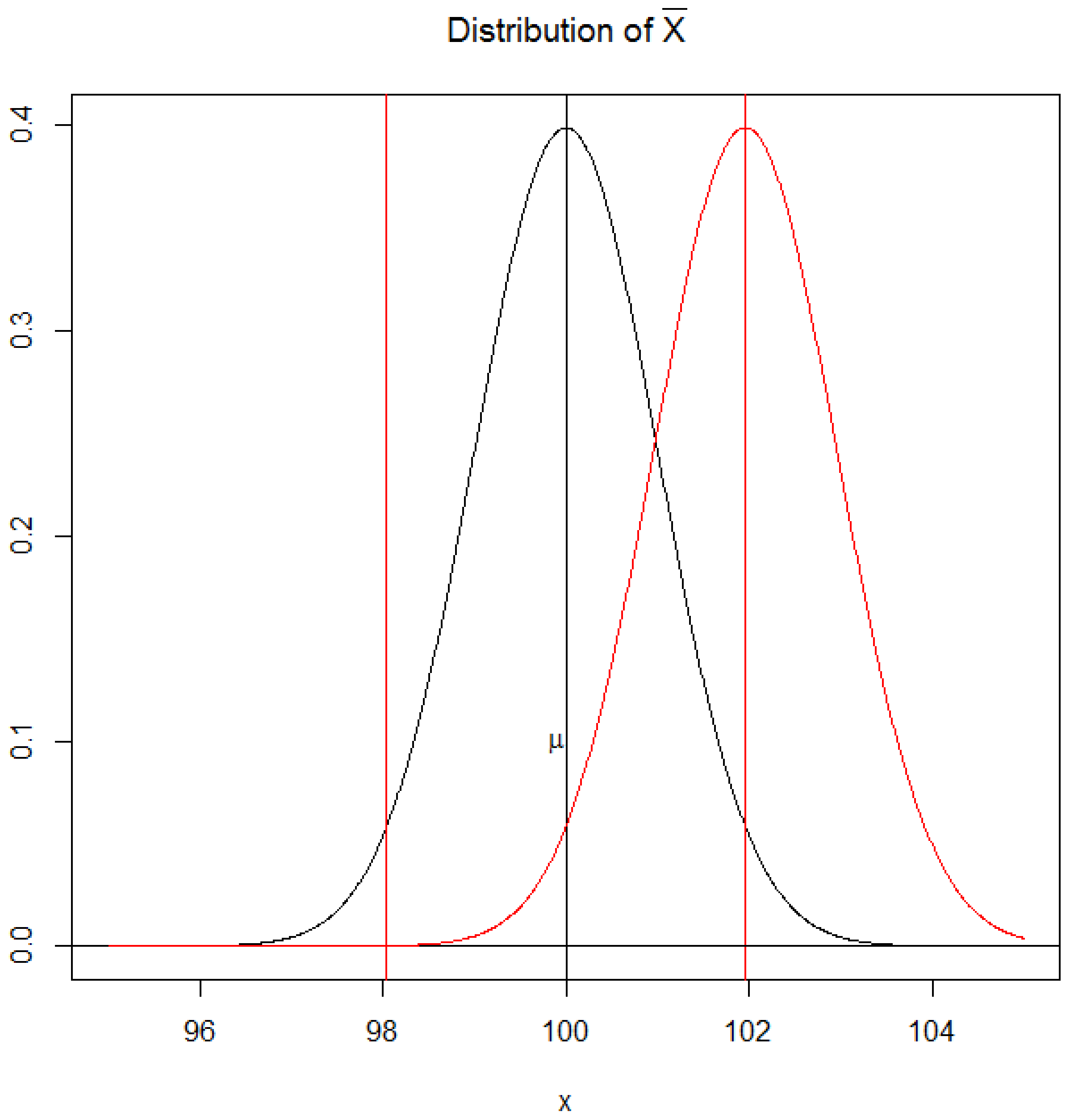
Compute $\Pr(X \leq x \mid p = p_0)$. If this is smaller than 0.025, then say that that value of p is not consistent. Note that when we are checking whether p_0 is consistent, we have

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

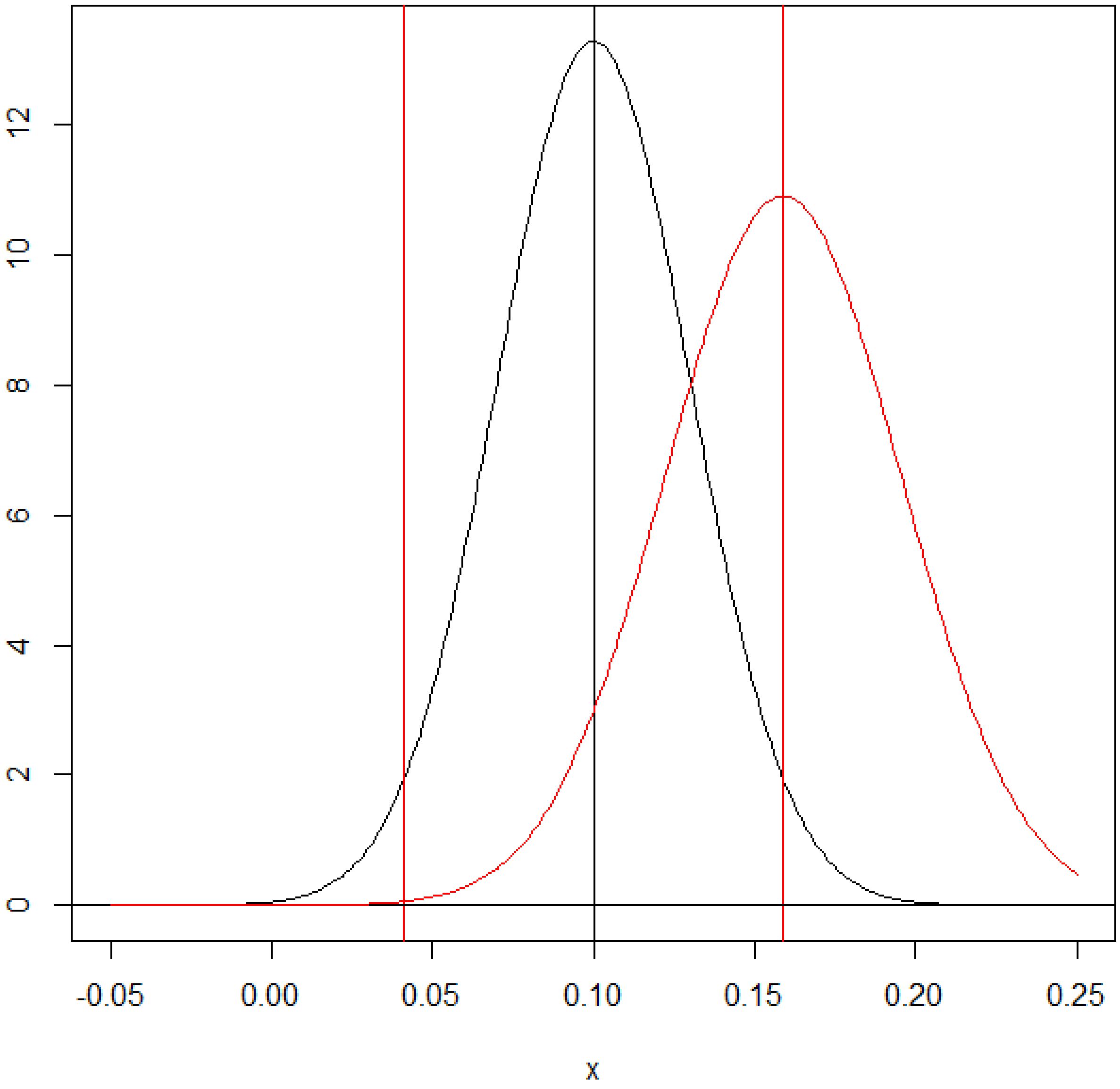
The value of p_0 that makes $\Pr(X \leq x \mid p = p_0) = 0.025$ forms the upper end of the 95% CI in this method, and similarly for the lower.

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


Distribution of X
0.00
0.01
0.02
0.03
0.04
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



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

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

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

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

Proportions

When making a confidence interval for a proportion,

we could use either

$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ or

$\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/n}$

where $\hat{p} = x/n$ and $\tilde{p} = (x+2)/(n+4)$

For a hypothesis test, we don't have this problem. The null hypothesis is

$p = p_0$ so under the null, we know what p is, we don't have to estimate it.

Under the null, the variance of $\hat{p}$ is $p(1-p)/n$, and under the null this is known.

Proportions

So if

$n = 600 \quad x = 217 \quad \hat{p} = 0.36167$

and under the null hypothesis that $p = 0.30$

$E(\hat{p}) = p = 0.30$

$V(\hat{p}) = p(1-p)/n = (0.30)(0.70)/600 = 0.00035$

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} = \frac{217/600-0.30}{\sqrt{(0.30)(0.70)/600}} = \frac{0.06167}{0.01871} = 3.296 = z$$

The p-value for this test is 2(0.00049) = 0.00098, so not consistent. We reject the null hypothesis.

We still need the normal approximation to make this correct, but there is no approximation to the variance.

Confidence Intervals and Hypothesis Tests

For means, the $100(1-\alpha)\%$ confidence interval is exactly the set of hypothetical values of μ that are not rejected by the usual hypothesis test if we set the cutoff for rejection at α. Otherwise put, a test of the hypothesis that $\mu = \mu_0$ is rejected at a significance level of α exactly when μ does not lie in the $100(1-\alpha)\%$ confidence interval for the mean.

This does not quite work for proportions because the ordinary confidence confidence interval uses a standard deviation of $\sqrt{\hat{p}(1-\hat{p})/n}$ to see if a particular value $p = p_0$ lies in the interval, whereas the test uses $\sqrt{p_0(1-p_0)/n}$ MATLAB's Clopper-Pearson interval (the default) does exactly invert the hypothesis test.

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.
- 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)

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
```

Binomial Hypothesis Tests in Matlab

Test stem cell data to see if 10% success has for sure been exceeded.

```
>> phat = 14/100
>> p0 = 0.10
>> sdp0 = sqrt(p0*(1-p0)/100)
>> [h p ci zval] = ztest(phat,p0,sdp0)
h =
     0               # not rejected
p =
    0.1824           # p-value
ci =
    0.0812           # this is not any of the usual CI choices
    0.1988           # because it uses p0 for the variance
zval =

    1.3333
```

Summary

For binomial data, if we have x successes out of n in a sample and if we wish to test $H_0 : p = p_0$, then we use the approximate z-statistic

$$z = \frac{x / n - p_0}{\sqrt{p_0(1 - p_0) / n}}$$

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 $\hat{p} = x / n$. For the Agresti-Coull interval favored by the book, we use $\tilde{p} = (x + 2) / (n + 4)$.

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 of Sample Size

To obtain a confidence interval with half-width w

$\hat{p} \pm w$

the required sample size n is

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

rounded up to the nearest whole number