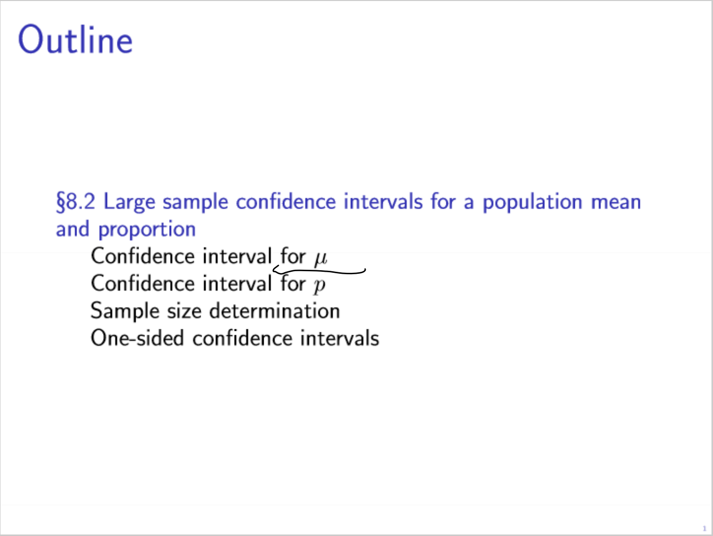
class11

February 5, 2021 10:20 AM

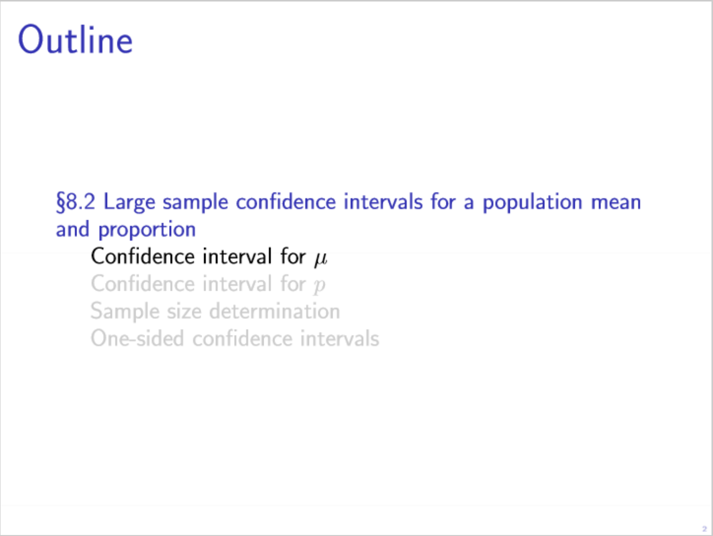


Outline

§8.2 Large sample confidence intervals for a population mean and proportion

- Confidence interval for μ
- Confidence interval for p
- Sample size determination
- One-sided confidence intervals

1

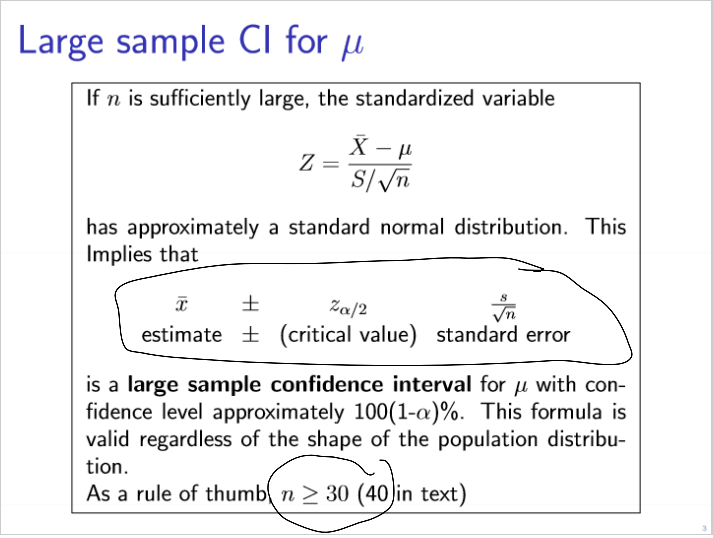


Outline

§8.2 Large sample confidence intervals for a population mean and proportion

- Confidence interval for μ
- Confidence interval for p
- Sample size determination
- One-sided confidence intervals

2



Large sample CI for μ

If n is sufficiently large, the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

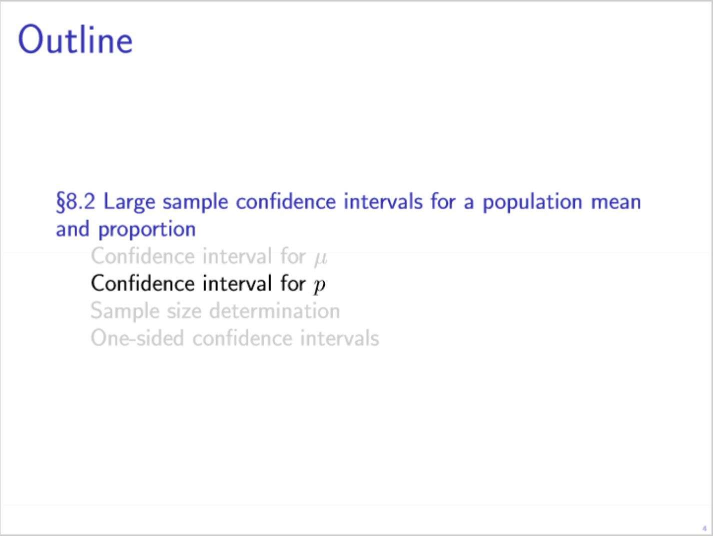
has approximately a standard normal distribution. This Implies that

$\bar{x}$	$\pm$	$z_{\alpha/2}$	$\frac{s}{\sqrt{n}}$
estimate	$\pm$	(critical value)	standard error

is a **large sample confidence interval** for μ with confidence level approximately 100(1-α)%. This formula is valid regardless of the shape of the population distribution.

As a rule of thumb $n \geq 30$ (40 in text)

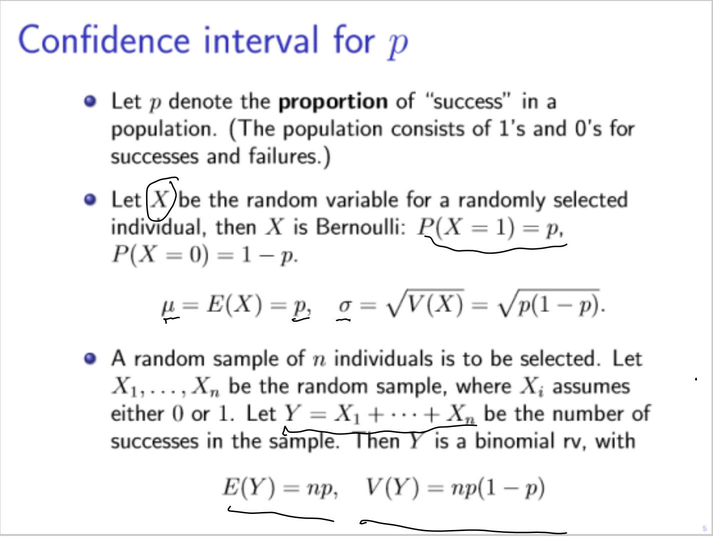
3



Outline

§8.2 Large sample confidence intervals for a population mean and proportion

- Confidence interval for μ
- Confidence interval for p
- Sample size determination
- One-sided confidence intervals



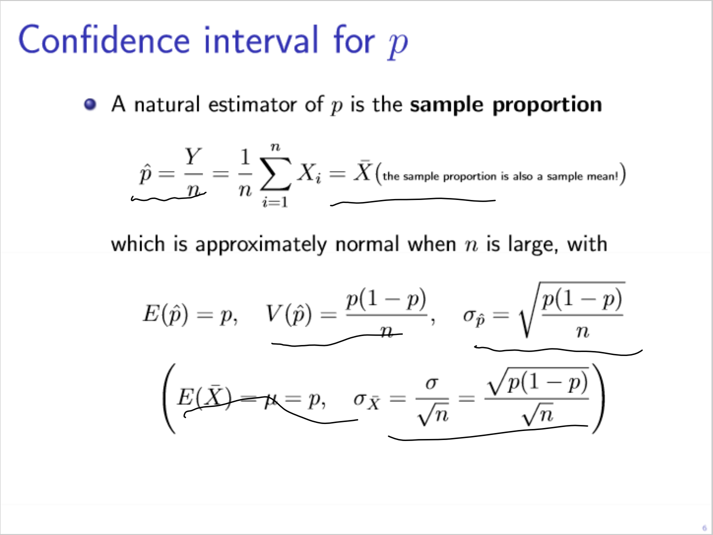
Confidence interval for p

- Let p denote the **proportion** of "success" in a population. (The population consists of 1's and 0's for successes and failures.)
- Let X be the random variable for a randomly selected individual, then X is Bernoulli: $P(X = 1) = p$, $P(X = 0) = 1 - p$.

$$\mu = E(X) = p, \quad \sigma = \sqrt{V(X)} = \sqrt{p(1-p)}.$$

- A random sample of n individuals is to be selected. Let $X_1, \ldots, X_n$ be the random sample, where X_i assumes either 0 or 1. Let $Y = X_1 + \cdots + X_n$ be the number of successes in the sample. Then Y is a binomial rv, with

$$E(Y) = np, \quad V(Y) = np(1-p)$$



Confidence interval for p

- A natural estimator of p is the **sample proportion**

$$\hat{p} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \left(\text{the sample proportion is also a sample mean!}\right)$$

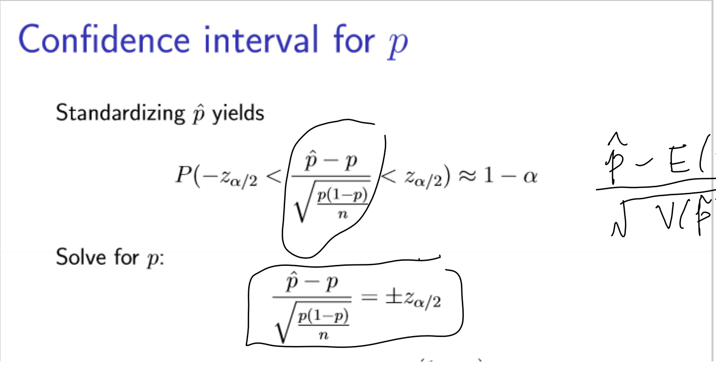
which is approximately normal when n is large, with

$$E(\hat{p}) = p, \quad V(\hat{p}) = \frac{p(1-p)}{n}, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\left(E(\bar{X}) = \mu = p, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right)$$

X standardise:

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0,1)$$



Confidence interval for p

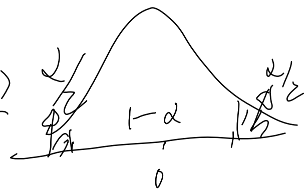
Standardizing $\hat{p}$ yields

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

Solve for p:

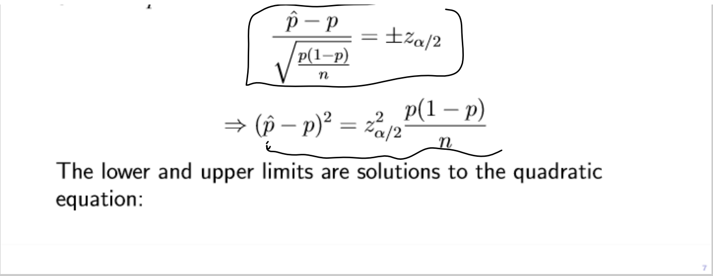
$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_{\alpha/2}$$

$$\frac{\hat{p} - E(\hat{p})}{\sqrt{V(\hat{p})}}$$



$$\sqrt{\frac{X-\mu}{\sigma}} \text{ or } \frac{X-\mu}{\sigma/\sqrt{n}}$$

$$\hat{p}^2 - 2\hat{p}\cdot p + p^2 = z_{\alpha/2}^2 \frac{p}{n} - z_{\alpha/2}^2 \frac{p^2}{n}$$



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

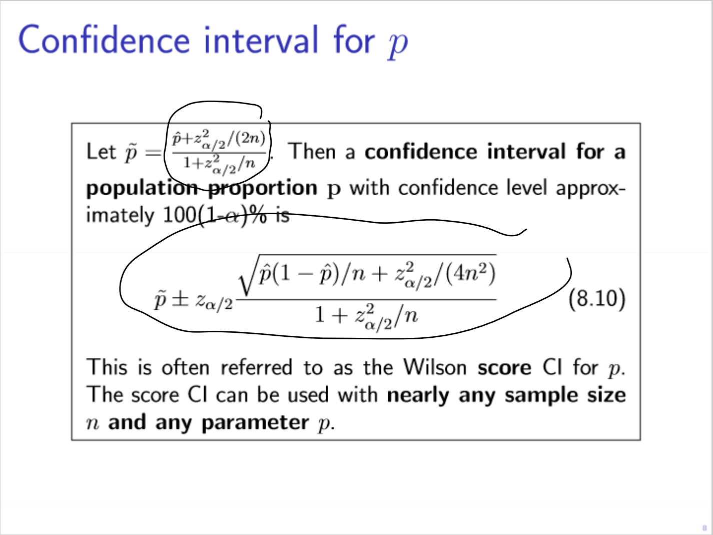
$$\Rightarrow (\hat{p}-p)^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}$$

The lower and upper limits are solutions to the quadratic equation:

$$\hat{p}^2 - 2\hat{p}\cdot p + p^2 = z_{\alpha/2}^2 \frac{p}{n} - z_{\alpha/2}^2 \frac{p^2}{n}$$

$$p^2\left(1+\frac{z_{\alpha/2}^2}{n}\right) + p\left(-2\hat{p} - z_{\alpha/2}^2\right) + \hat{p}^2 = 0$$

$$ap^2 + bp + c = 0$$

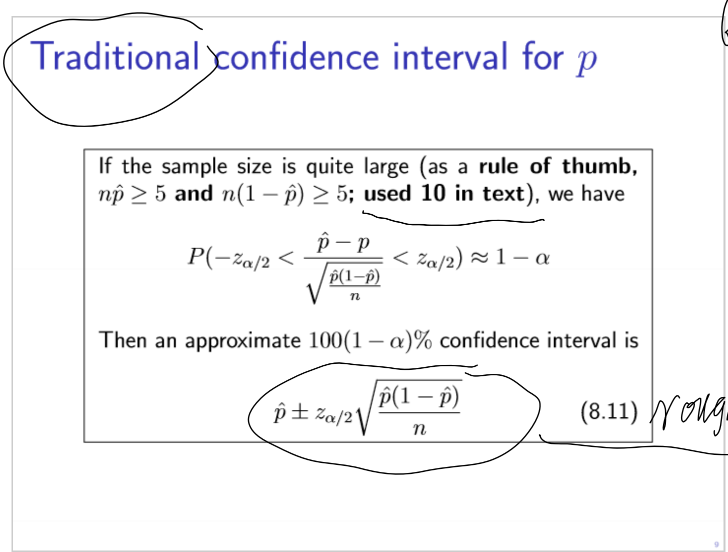


Confidence interval for p

Let $\tilde{p} = \frac{\hat{p}+z_{\alpha/2}^2/(2n)}{1+z_{\alpha/2}^2/n}$. Then a **confidence interval for a population proportion** p with confidence level approximately 100(1-α)% is

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1+z_{\alpha/2}^2/n} \qquad (8.10)$$

This is often referred to as the Wilson **score** CI for p. The score CI can be used with **nearly any sample size** n **and any parameter** p.



Traditional confidence interval for p

General $n \geq 30$ for CI for μ

If the sample size is quite large (as a **rule of thumb,** $n\hat{p} \geq 5$ **and** $n(1-\hat{p}) \geq 5$**; used 10 in text**), we have

$n\hat{p} = \#$ successes ≥ 5

$n(1-\hat{p}) = n - n\hat{p} = \#$ failures ≥ 5.

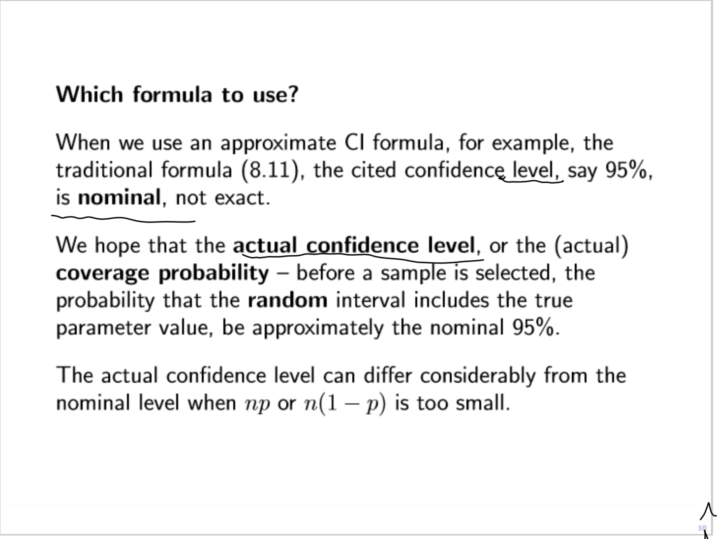
$$P(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}) \approx 1-\alpha$$

Then an approximate $100(1-\alpha)\%$ confidence interval is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (8.11)$$

roughly

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



Which formula to use?

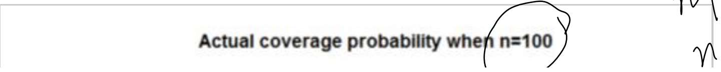
When we use an approximate CI formula, for example, the traditional formula (8.11), the cited confidence level, say 95%, is **nominal**, not exact.

We hope that the **actual confidence level**, or the (actual) **coverage probability** – before a sample is selected, the probability that the **random** interval includes the true parameter value, be approximately the nominal 95%.

The actual confidence level can differ considerably from the nominal level when np or $n(1-p)$ is too small.

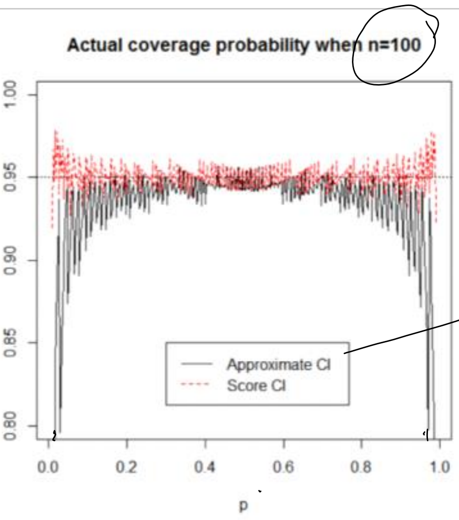
$n\hat{p} \geq 5$

$n(1-\hat{p}) \geq 5$

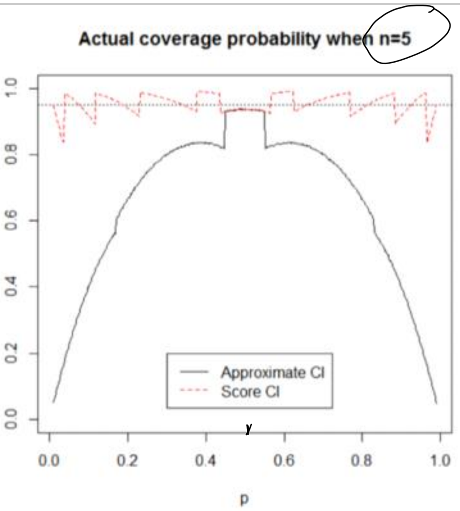


Actual coverage probability when n=100

$n\hat{p} \geq$
$n(1-\hat{p}) \geq 5$



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



The text recommends that the score CI (8.10) always be used.

We will use in the course the traditional formula (8.11).

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

success . allergic to sulfites

n

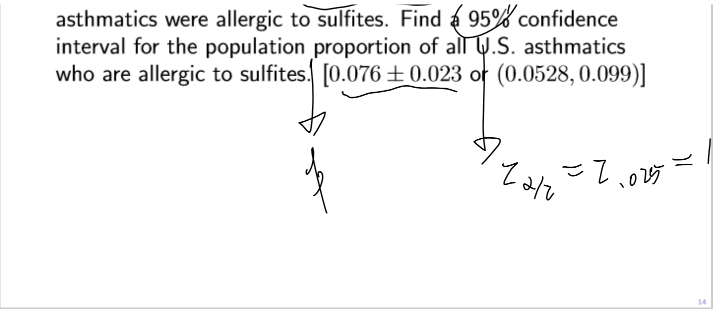
Example. Studies are performed to estimate the percentage of the nation's 10 million asthmatics who are allergic to sulfites. In one survey, 38 of 500 randomly selected U.S. asthmatics were allergic to sulfites. Find a 95% confidence interval for the population proportion of all U.S. asthmatics who are allergic to sulfites. $[0.076 \pm 0.023$ or $(0.0528, 0.099)]$

$\hat{p} = \frac{38}{500}$. Is n large enough?

$n\hat{p} = 38 \geq 5$

$n - n\hat{p} = 500 - 38 \geq 5$

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



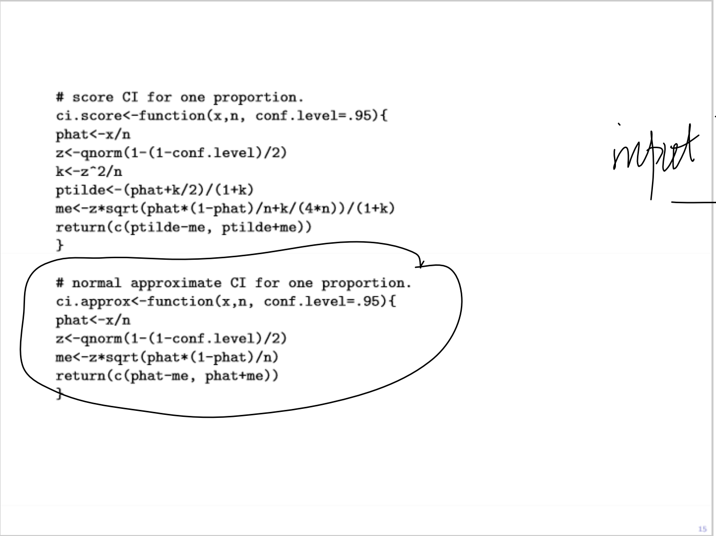
asthmatics were allergic to sulfites. Find a 95% confidence interval for the population proportion of all U.S. asthmatics who are allergic to sulfites. [0.076 ± 0.023 or $(0.0528, 0.099)$]

$\hat{p}$

$z_{\alpha/2} = z_{.025} = 1.96$

$n - np = 500 - 38 > 5$

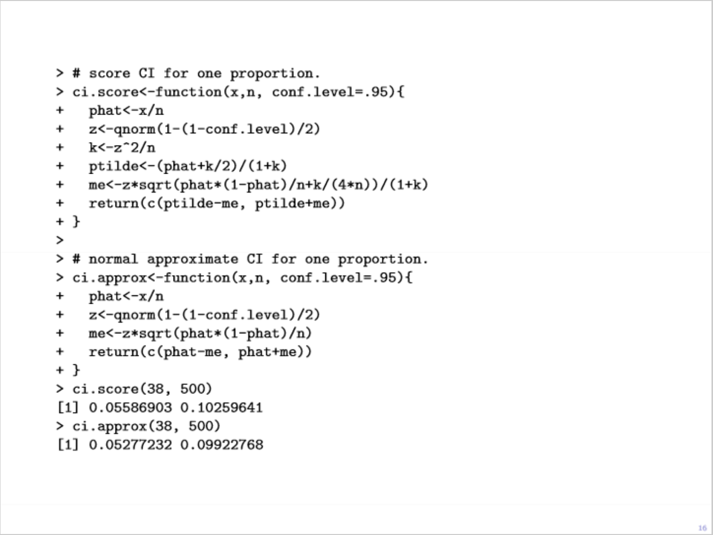
$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$



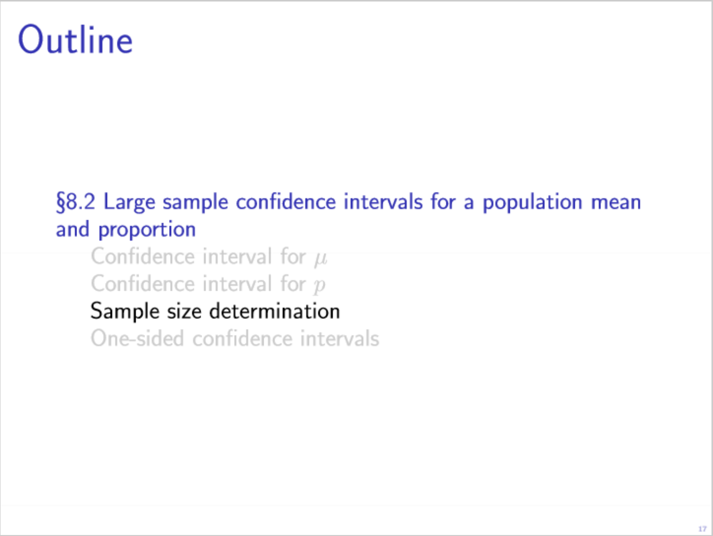
```
# score CI for one proportion.
ci.score<-function(x,n, conf.level=.95){
phat<-x/n
z<-qnorm(1-(1-conf.level)/2)
k<-z^2/n
ptilde<-(phat+k/2)/(1+k)
me<-z*sqrt(phat*(1-phat)/n+k/(4*n))/(1+k)
return(c(ptilde-me, ptilde+me))
}

# normal approximate CI for one proportion.
ci.approx<-function(x,n, conf.level=.95){
phat<-x/n
z<-qnorm(1-(1-conf.level)/2)
me<-z*sqrt(phat*(1-phat)/n)
return(c(phat-me, phat+me))
}
```

input $x, n, 1-\alpha$ $\rightarrow$ $\boxed{\frac{x}{n} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}}$ $\rightarrow$



```
> # score CI for one proportion.
> ci.score<-function(x,n, conf.level=.95){
+   phat<-x/n
+   z<-qnorm(1-(1-conf.level)/2)
+   k<-z^2/n
+   ptilde<-(phat+k/2)/(1+k)
+   me<-z*sqrt(phat*(1-phat)/n+k/(4*n))/(1+k)
+   return(c(ptilde-me, ptilde+me))
+ }
>
> # normal approximate CI for one proportion.
> ci.approx<-function(x,n, conf.level=.95){
+   phat<-x/n
+   z<-qnorm(1-(1-conf.level)/2)
+   me<-z*sqrt(phat*(1-phat)/n)
+   return(c(phat-me, phat+me))
+ }
> ci.score(38, 500)
[1] 0.05586903 0.10259641
> ci.approx(38, 500)
[1] 0.05277232 0.09922768
```



Outline

§8.2 Large sample confidence intervals for a population mean and proportion

- Confidence interval for μ
- Confidence interval for p
- Sample size determination
- One-sided confidence intervals

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ← margin of error

$$\hat p \pm z_{\alpha/2}\sqrt{\frac{\hat p(1-\hat p)}{n}}$$ ← margin of error

Sample size determination: Given a confidence level, say 95% and a margin of error (ME), say $r = .2$, how large should the sample size be?

The large sample z interval for μ:

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

$$z_{\alpha/2}\frac{s_0}{\sqrt n} \le r = .2$$

Need a **prior** and/or rough estimate of s, say s_0. Then we may solve the equation

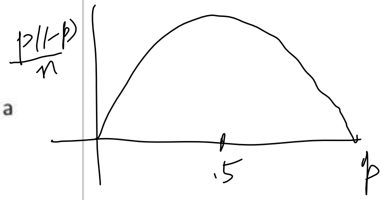
$$z_{\alpha/2}\frac{s_0}{\sqrt n} \le r \Rightarrow \boxed{n \ge \frac{z^2_{\alpha/2}s_0^2}{r^2}}$$

The traditional z interval for p:

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

1. If there is some prior information about the true proportion, for example from a pilot study, say $p \approx p_0$.

$$z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} = r \Rightarrow \boxed{n \ge \frac{z^2_{\alpha/2}p_0(1-p_0)}{r^2}}$$



2. If there is no prior information, use $p_0 = 0.5$, which is a "worst-case scenario".

$$\boxed{n \ge \frac{z^2_{\alpha/2}(0.5)(1-0.5)}{r^2} = \frac{z^2_{\alpha/2}}{4r^2}}$$

In the "allergic to sulfites" example, what sample size of another study is necessary such that the 95% CI for the true proportion has a margin of error at most 0.03 (or equivalently the width of the CI is at most 0.06)?

1. Use the information in the example, $\hat p = .076 \Rightarrow p_0 = 0.076$.

$$\hat p = \frac{38}{500} = .076$$

$$1.96\sqrt{\frac{(.076)(1-.076)}{n}} = .03$$

$$n \ge \cdots = 299.747 \approx 300.$$

2. Without prior estimate of p.

$$n \ge \cdots = 1067.111 \approx 1068$$

$$1.96\sqrt{\frac{(.5)(1-.5)}{n}} = .03$$