4.5 Confidence Intervals for a Proportion

Let Z be $N(0,1)$ and p be a number between 0 and 1; *critical z-value* z_p is

$$P(Z > z_p) = 1 - \Phi(z_p) = p.$$

Let $0 < \alpha < 1$ and x be number of successes in n observed trials of a Bernoulli experiment with unknown probability of success p. For $\hat{p} = \frac{x}{n}$, the $100(1-\alpha)\%$ confidence interval for proportion p is

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

where

$$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \text{and} \quad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

are the *margin of error* and *standard deviation* of the proportion respectively and α is the *level of significance*. We assume a large random sample is chosen, both $np \geq 5$ and $np(1-p) \geq 5$ and the conditions of a binomial distribution is satisfied. Also, *one-sided confidence interval estimates* for p include lower and upper bound respectively:

$$\left[\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right], \quad \left[0, \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Exercise 4.5 (Confidence Intervals for a Proportion)

1. *Confidence interval (CI) for proportion, p, of purchase slips made with Visa.* It is found 54 of 180 (or $\hat{p} = \frac{54}{180} = 0.3$) randomly selected from all credit card purchase slips are made with Visa where conditions of binomial distribution are satisfied. Calculate a 95% confidence interval (CI) of proportion p of purchase slips made with Visa.

   (a) *Point estimate.*
   Point estimate of *population* (actual, true) proportion of *all* credit card purchase slips made with Visa, p, is
   $\hat{p} =$ (i) **0.3** (ii) **54** (iii) **180**.
   Statistic $\hat{p} = 0.3$ probably does not exactly equal unknown parameter p.

   (b) *Check assumptions.*
   Since random sample chosen,
   conditions of binomial distribution are satisfied,
   and $np(1-p) \approx n\hat{p}(1-\hat{p}) = 180(0.3)(0.7) = 37.8 \geq 5$,
   and $np \approx n\hat{p} = 180(0.3) = 54 \geq 5$,
   assumptions (i) **have** (ii) **have not** been satisfied
   and so it is appropriate $\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ estimate parameter p.

(c) *95% Confidence Interval (CI) using R.*
The 95% CI for proportion of all credit cards made with Visa, p, is
(i) $\mathbf{(0.251, 0.349)}$ (ii) $\mathbf{(0.273, 0.367)}$ (iii) $\mathbf{(0.233, 0.367)}$.

```
prop1.interval <- function(x,n,conf.level) # function of 1-proportion CI for p
{
  p <- x/n
  z.crit <- -1*qnorm((1-conf.level)/2)
  margin.error <- z.crit*sqrt(p*(1-p)/n)
  ci.lower <- p - margin.error
  ci.upper <- p + margin.error
  dat <- c(p, z.crit, margin.error, ci.lower, ci.upper)
  names(dat) <- c("Mean", "Critical Value", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
prop1.interval(54,180,0.95) # 1-proportion 95% CI for p
```

```
           Mean  Critical Value Margin of Error        CI lower        CI upper
     0.30000000      1.95996398      0.06694551      0.23305449      0.36694551
```

where this *interval* includes not only smallest possible proportion of 0.233 and largest possible proportion of 0.367, but also other proportions in between these two extremes such as point estimate, $\hat{p} = 0.3$.
Length of this CI is $L \approx 0.367 - 0.233 = 0.134$.
So, 95% confident population parameter p in (0.233, 0.367).

(d) *90% CI using R.*
The 90% CI for proportion of all credit cards made with Visa, p, is
(i) $\mathbf{(0.251, 0.349)}$ (ii) $\mathbf{(0.244, 0.356)}$ (iii) $\mathbf{(0.233, 0.367)}$.
Length of this CI is $L \approx 0.356 - 0.244 = 0.112$.

```
prop1.interval(54,180,0.90) # 1-proportion 90% CI for p
```

```
           Mean  Critical Value Margin of Error        CI lower        CI upper
     0.30000000      1.64485363      0.05618245      0.24381755      0.35618245
```

(e) *85% CI using R.*
The 85% CI for proportion of all credit cards made with Visa, p, is
(i) $\mathbf{(0.251, 0.349)}$ (ii) $\mathbf{(0.273, 0.367)}$ (iii) $\mathbf{(0.233, 0.367)}$.
Length of this CI is $L \approx 0.349 - 0.251 = 0.098$.

```
prop1.interval(54,180,0.85) # 1-proportion 85% CI for p
```

```
           Mean  Critical Value Margin of Error        CI lower        CI upper
     0.30000000      1.43953147      0.04916936      0.25083064      0.34916936
```

(f) *Comparing CI lengths.*
Length of 95% CI for p, $L = 0.134$, is
(i) **longer than** (ii) **same length as** (iii) **shorter than**
length of 90% CI for p, $L = 0.112$, which is
(i) **longer than** (ii) **same length as** (iii) **shorter than**
length of 85% CI for p, $L = 0.098$.
Increasing confidence increases CI length.

(g) *Margin of error.*
Half of length, L, is margin of error, $E = \frac{L}{2}$.
Consequently, for 95% CI for p,

$E = \frac{L}{2} = \frac{0.134}{2} =$ (i) **0.067** (ii) **0.056** (iii) **0.049**,
and for 90% CI for p,
$E = \frac{L}{2} = \frac{0.112}{2} =$ (i) **0.067** (ii) **0.056** (iii) **0.049**,
and for 85% CI for p,
$E = \frac{L}{2} = \frac{0.098}{2} =$ (i) **0.067** (ii) **0.056** (iii) **0.049**,

(h) *Other ways of writing confidence intervals.*
Different possible ways of writing 95% CI include (choose *one or more*!)

i. $\mathbf{(0.233, 0.367)}$
ii. $\mathbf{(0.3 - 0.067, 0.3 + 0.067)}$
iii. $\mathbf{0.3 \pm 0.067}$

where $\hat{p} = 0.3$ is *point estimate* and $E = 0.067$ is margin of error.
In a similar way,
90% CI of parameter p is
(i) $\mathbf{0.3 \pm 0.067}$ (ii) $\mathbf{0.3 \pm 0.056}$ (iii) $\mathbf{0.3 \pm 0.049}$
and 85% CI of parameter p is
(i) $\mathbf{0.3 \pm 0.067}$ (ii) $\mathbf{0.3 \pm 0.056}$ (iii) $\mathbf{0.3 \pm 0.049}$

(i) *CI using formula, a first look.*
Since $\hat{p} = 0.3$ and $n = 180$, the 95% CI for p is

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

the *incomplete* answer

i. $\mathbf{180 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$,
ii. $\mathbf{0.3 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$,
iii. $\mathbf{0.7 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.7(1-0.3)}{180}}}$.

In a similar way,
90% CI is
(i) $\mathbf{0.3 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$ (ii) $\mathbf{0.7 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$ (iii) $\mathbf{180 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$
and 85% CI is
(i) $\mathbf{0.3 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$ (ii) $\mathbf{0.7 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$ (iii) $\mathbf{180 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.3(1-0.3)}{180}}}$
All three CIs same except three critical values, $z_{\frac{\alpha}{2}}$, are different.

(j) *CI using formula: calculating critical value, $z_{\frac{\alpha}{2}}$.*
Critical value for 95% $= (1-\alpha) \cdot 100\% = (1 - 0.05) \cdot 100\%$ CI is
$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} =$ (i) **1.96** (ii) **1.645** (iii) **1.44**.

```
qnorm(0.975) # critical value z_0.05/2
```

```
> qnorm(0.975) # critical value z_0.05/2
[1] 1.959964
```

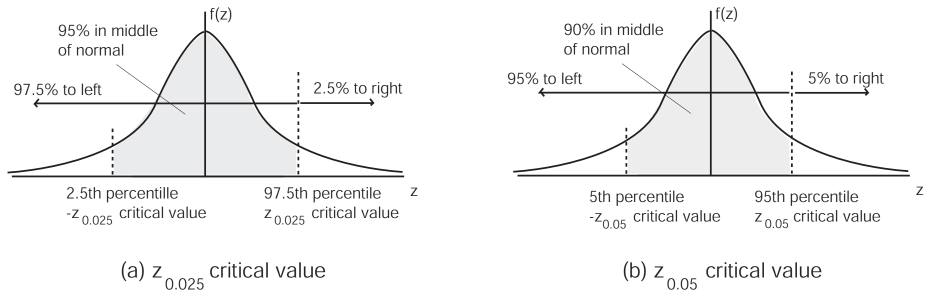

Figure 4.5: Critical values

Critical value for 90% $= (1-\alpha)\cdot 100\% = (1-0.10)\cdot 100\%$ CI is $z_{\frac{\alpha}{2}} = z_{\frac{0.10}{2}} = z_{0.05} =$ (i) **1.96** (ii) **1.645** (iii) **1.44**.

```
qnorm(0.95) # critical value z_0.1/2
```

```
> qnorm(0.95) # critical value z_0.1/2
[1] 1.644854
```

Critical value for 85% $= (1-\alpha)\cdot 100\% = (1-0.15)\cdot 100\%$ CI is $z_{\frac{\alpha}{2}} = z_{\frac{0.15}{2}} = z_{0.075} =$ (i) **1.96** (ii) **1.645** (iii) **1.44**.

```
qnorm(0.925) # critical value z_0.15/2
```

```
> qnorm(0.925) # critical value z_0.15/2
[1] 1.439531
```

(k) *CI using formula.*
A 95% CI for proportion of Visa credit card purchase slips, p, is $\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} =$

i. $\mathbf{0.3 \pm 1.96 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$
ii. $\mathbf{0.3 \pm 1.645 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$
iii. $\mathbf{0.3 \pm 1.44 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$

and a 90% CI for proportion of Visa credit card purchase slips, p, is

i. $\mathbf{0.3 \pm 1.96 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$
ii. $\mathbf{0.3 \pm 1.645 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$
iii. $\mathbf{0.3 \pm 1.44 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$

and an 85% CI for proportion of Visa credit card purchase slips, p, is

i. $\mathbf{0.3 \pm 1.96 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$
ii. $\mathbf{0.3 \pm 1.645 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$

iii. $\mathbf{0.3 \pm 1.44 \times \sqrt{\frac{0.3(1-0.3)}{180}}}$

(l) *Population, Sample, Statistic and Parameter.* Match columns.

terms	credit card example
(a) population	**(a)** Visa or not, all purchase slips
(b) sample	**(b)** proportion of all slips made with Visa, p
(c) statistic	**(c)** Visa or not, 180 purchase slips
(d) parameter	**(d)** proportion of 180 slips made with Visa, $\hat{p}$

terms	(a)	(b)	(c)	(d)
credit card example				

2. *95% CI, proportion of student heights over 6 feet tall.*
37 of 102 students, chosen at random from PNW, over 6 feet tall.

(a) *Point estimate*
Point estimate of proportion, p, of student heights over 6 feet tall is
$\hat{p} = \frac{37}{102} \approx$ (i) **0.363** (ii) **0.378** (iii) **0.391**.

(b) *Check assumptions.*
Since $np \approx n\hat{p} = 102\left(\frac{37}{102}\right) = 37 \geq 5$,
and $np(1-p) \approx n\hat{p}(1-\hat{p}) = 102\left(\frac{37}{102}\right)\left(1 - \frac{37}{102}\right) \approx 23.6 > 5$,
assumptions (i) **have** (ii) **have not** been satisfied
and so it is appropriate $\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ estimate parameter p.

(c) *Using R.* The 95% CI for p is
(i) $\mathbf{(0.269, 0.456)}$ (ii) $\mathbf{(0.273, 0.367)}$ (iii) $\mathbf{(0.233, 0.367)}$.

```
prop1.interval(37,102,0.95) # 1-proportion 95% CI for p

          Mean  Critical Value Margin of Error        CI lower        CI upper
     0.3627451       1.9599640       0.0933051       0.2694400       0.4560502
```

(d) *Using formula: critical value using R.*
Critical value for 95% $= (1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI for p is
$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} =$ (i) **1.28** (ii) **1.96** (iii) **2.58**.

```
qnorm(0.975) # critical value z_0.05/2 for 95% CI

> qnorm(0.975) # critical value z_0.05/2
[1] 1.959964
```

(e) *Using formula: critical value using Table C.1.*
Critical value for 95% $= (1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI for p is
$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} =$ (i) **1.28** (ii) **1.96** (iii) **2.58**.

(f) *Using formula.*
Since $\hat{p} = \frac{37}{102}$ and $n = 102$, the 95% CI for p is
$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} =$

(i) $\mathbf{0.36 \pm 1.28 \times \sqrt{\frac{0.36(1-0.36)}{102}}}$
(ii) $\mathbf{0.36 \pm 1.96 \times \sqrt{\frac{0.36(1-0.36)}{102}}}$
(iii) $\mathbf{0.36 \pm 2.58 \times \sqrt{\frac{0.36(1-0.36)}{102}}}$
$\approx (0.269, 0.456)$

(g) *Length*, L, of 95% CI is
$L = 0.456 - 0.269 =$ (i) **0.176** (ii) **0.187** (iii) **0.354**.
Half of length, margin of error,
$E = \frac{L}{2} =$ (i) **0.088** (ii) **0.0935** (iii) **0.177**.
Notice, margin of error also equals

$$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times \sqrt{\frac{\frac{37}{102}(1-\frac{37}{102})}{102}} \approx 0.0935.$$

(h) *Confidence Level and Sample Size.*
The larger the confidence level (critical value, $z_{\frac{\alpha}{2}}$)
the (i) **larger** (ii) **smaller** the margin of error.
The larger the sample size, n, the
(i) **larger** (ii) **smaller** the margin of error.

(i) *One-sided confidence level with upper bound using R.*
The 95% CI for p with upper bound is

$$\left(0, \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) =$$

(i) $\mathbf{(0.269, 0.456)}$ (ii) $\mathbf{(0.273, 1)}$ (iii) $\mathbf{(0, 0.441)}$.

```
prop1.interval <- function(x,n,conf.level) # function of 1-proportion CI with upper bound for p
{
  p <- x/n
  z.crit <- -1*qnorm(1-conf.level)
  margin.error <- z.crit*sqrt(p*(1-p)/n)
  ci.lower <- 0
  ci.upper <- p + margin.error
  dat <- c(p, z.crit, margin.error, ci.lower, ci.upper)
  names(dat) <- c("Mean", "Critical Value", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
prop1.interval(37,102,0.95) # 1-proportion 95% CI with upper bound for p

          Mean  Critical Value Margin of Error        CI lower        CI upper
    0.36274510      1.64485363      0.07830411      0.00000000      0.44104920
```

4.6 Confidence Intervals for a Mean

Let $\bar{x}$ be the mean with sample of size n taken from a population with *know* variance σ^2 and unknown mean μ and $0 < \alpha < 1$. The $(1-\alpha) \cdot 100\%$ confidence interval for

μ is called a *z-interval*:

$$\bar{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right).$$

The $(1-\alpha) \cdot 100\%$ confidence interval for μ with *un*known σ is called a *t-interval*:

$$\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right),$$

where $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ has a Student-t distribution and where

$$E = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) \quad \text{and} \quad \left(\frac{s}{\sqrt{n}} \right)$$

are the *margin of error* and *standard error* of the mean respectively and α is the *level of significance*. We assume a large random sample, where either the underlying distribution is normal with no outliers or if the sample size large ($n > 30$). Also, *one-sided confidence interval estimates* for μ include lower and upper bound respectively:

$$\left(\bar{x} - t_\alpha \left(\frac{s}{\sqrt{n}} \right), \infty \right), \quad \left(-\infty, \bar{x} + t_\alpha \left(\frac{s}{\sqrt{n}} \right) \right).$$

Exercise 4.6 (Confidence Intervals for a Mean)

1. *Estimates for population average weight of PNW students.*
   Average weight of simple random sample of 11 PNW students is $\bar{x} = 167$ pounds with sample SD $s = 20.1$ pounds. Weights normally distributed, no outliers.

   (a) *Point estimate.*
   Point estimate of population weight of *all* students, μ, is
   $\bar{x} =$ (i) **11** (ii) **20.1** (iii) **167**.
   Also notice σ is *un*known and *estimated* by $s = 20.1$.

   (b) *95% CI*

   i. *Using R.* The 95% CI for μ is
   (i) $\mathbf{(143.5, 182.5)}$ (ii) $\mathbf{(151.5, 180.5)}$ (iii) $\mathbf{(153.5, 180.5)}$.

```
mean1.t.interval <- function(m,s,n,conf.level)
{
  t.crit <- -1*qt((1-conf.level)/2,n-1)
  margin.error <- t.crit*s/sqrt(n)
  ci.lower <- m - margin.error
  ci.upper <- m + margin.error
  dat <- c(mean, t.crit, margin.error, ci.lower, ci.upper)
  names(dat) <- c("Mean", "Critical Value", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
mean1.t.interval(167,20.1,11,0.95) # m: mean, s: SD, n: sample size, 95% t-interval
```

```
            Mean  Critical Value Margin of Error        CI lower        CI upper
      167.000000        2.228139       13.503364      153.496636      180.503364
```

So, 95% confident population parameter μ in (153.5, 180.5).

ii. *Using formula: degrees of freedom (df).*
df $= n - 1 = 11 - 1 =$ (i) **10** (ii) **11**.

iii. *Using formula: critical value using R.*
Critical value 95% $= (1-\alpha) \cdot 100\% = (1-0.05) \cdot 100\%$ CI, 10 df
$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \approx$ (i) **1.28** (ii) **2.23** (iii) **2.58**.

```
qt(0.975,10) # critical value t, 10 df, for 95% CI
```

```
> qt(0.975,10) # critical value t for 95% CI
[1] 2.228139
```

iv. *Using formula: critical value using Table C.3.*
Critical value 95% $= (1-\alpha) \cdot 100\% = (1-0.05) \cdot 100\%$ CI, 10 df
$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \approx$ (i) **1.28** (ii) **2.23** (iii) **2.58**.

v. *Using formula.*
The 95% CI for μ is
$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$
(i) $\mathbf{20.1 \pm 167 \times \frac{2.23}{\sqrt{11}}}$ (ii) $\mathbf{2.23 \pm 167 \times \frac{20.1}{\sqrt{11}}}$ (iii) $\mathbf{167 \pm 2.23 \times \frac{20.1}{\sqrt{11}}}$
which equals
(i) $\mathbf{20.1 \pm 12.51}$ (ii) $\mathbf{2.23 \pm 13.51}$ (iii) $\mathbf{167 \pm 13.51} \approx (153.5, 180.5)$.

(c) *99% CI*

i. *Using R.* The 99% CI for μ is
(i) $\mathbf{(147.8, 186.2)}$ (ii) $\mathbf{(151.5, 180.5)}$ (iii) $\mathbf{(153.5, 180.5)}$.

```
mean1.t.interval(167,20.1,11,0.99) # m: mean, s: SD, n: sample size, 99% t-interval
```

```
            Mean  Critical Value Margin of Error        CI lower        CI upper
      167.000000        3.169273       19.206990      147.793010      186.206990
```

So, 99% confident population parameter μ in (147.8, 186.2).

ii. *Using formula: degrees of freedom.*
df $= n - 1 = 11 - 1 =$ (i) **10** (ii) **11**.

iii. *Using formula: critical value.*
Critical value 99% $= (1-\alpha) \cdot 100\% = (1-0.01) \cdot 100\%$ CI, 10 df
$t_{\frac{\alpha}{2}} = t_{\frac{0.01}{2}} = t_{0.005} \approx$ (i) **1.28** (ii) **2.23** (iii) **3.17**.

```
qt(0.995,10) # critical value t, 10 df, for 99% CI
```

```
[1] 3.169273
```

iv. *Using formula.*
The 99% CI for μ is
$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$
(i) $\mathbf{20.1 \pm 20.1 \times \frac{3.17}{\sqrt{11}}}$ (ii) $\mathbf{3.17 \pm 167 \times \frac{20.1}{\sqrt{11}}}$ (iii) $\mathbf{167 \pm 3.17 \times \frac{20.1}{\sqrt{11}}}$.
which equals
(i) $\mathbf{20.1 \pm 19.21}$ (ii) $\mathbf{3.17 \pm 19.21}$ (iii) $\mathbf{167 \pm 19.21} \approx (147.8, 186.2)$

v. *Margin of error if σ approximated by $s \approx 20.1$.*

$$E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \approx 3.17 \times \frac{20.1}{\sqrt{11}} \approx$$

(i) **11** (ii) **15.6** (iii) **19.2**.

vi. *Margin of error if $\sigma = 20.1$ known.*

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \approx 2.58 \times \frac{20.1}{\sqrt{11}} \approx$$

(i) **11** (ii) **15.6** (iii) **19.2**.

```
qnorm(0.995) # critical value z for 99% CI
```

```
[1] 2.575829
```

vii. (i) **True** (ii) **False** There is a 99% *chance* population average weight, μ, falls in sample interval (147.8, 186.2).

2. *Confidence interval for average length of corn cobs, raw data.*
Corn cob lengths for $n = 15 < 30$ cobs, chosen at random, are noted.

$$18, 23, 24, 20, 21, 19, 27, 24, 19, 20, 25, 20, 18, 26, 20$$

(a) *Point estimate.*
Point estimate of population length of *all* cobs, μ, is
$\bar{x} =$ (i) **2.97** (ii) **21.6** (iii) **15**.
Also notice population SD in cob length, σ, is *un*known and *estimated* by
$s \approx$ (i) **2.97** (ii) **21.6** (iii) **15**.

```
corn <- c(18,23,24,20,21,19,27,24,19,20,25,20,18,26,20)
m <- mean(corn); m; s <- sqrt(var(corn)); s; n <- length(corn); n
```

```
[1] 21.6
[1] 2.971291
[1] 15
```

(b) *95% CI*

i. *Using R.* The 95% CI for μ is
(i) **(17.96, 21.24)** (ii) **(19.96, 22.24)** (iii) **(19.96, 23.25)**.

```
mean1.t.interval(m,s,n,0.95) # m: mean, s: SD, n: sample size, 95% t-interval
```

```
       Mean  Critical Value Margin of Error       CI lower       CI upper
  21.600000       2.144787        1.645446      19.954554      23.245446
```

ii. *Using formula: degrees of freedom (df).*
df $= n - 1 =$ (i) **15** (ii) **14**.

iii. *Using formula: critical value.*
Critical value 95% $= (1 - \alpha) \cdot 100\% = (1 - 0.05) \cdot 100\%$ CI, 14 df
$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \approx$ (i) **1.76** (ii) **2.15**.

```
qt(0.975,14) # critical value t, 14 df, for 99% CI
```

```
[1] 2.144787
```

iv. *Using formula.*
The 95% CI for μ is
$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$
(i) $\mathbf{21.6 \pm 2.15 \times \frac{2.97}{\sqrt{15}}}$ (ii) $\mathbf{21.6 \pm 2.15 \times \frac{3.97}{\sqrt{15}}}$ (iii) $\mathbf{21.6 \pm 3.15 \times \frac{2.97}{\sqrt{15}}}$.

(c) *99% CI*

i. *Using R.* The 99% CI for μ is
(i) $\mathbf{(19.23, 23.45)}$ (ii) $\mathbf{(19.96, 23.24)}$ (iii) $\mathbf{(19.32, 23.88)}$.

```
mean1.t.interval(m,s,n,0.99) # m: mean, s: SD, n: sample size, 99% t-interval
```

```
         Mean  Critical Value Margin of Error        CI lower        CI upper
    21.600000        2.976843        2.283786       19.316214       23.883786
```

ii. *Using formula: degrees of freedom (df).*
The df, here, for 99% CI is (i) **same as** (ii) **different from** degrees of freedom calculated for 95% CI above because same sample size is used in both cases.

iii. *Using formula: critical value.*
Critical value 99% $= (1 - \alpha) \cdot 100\% = (1 - 0.01) \cdot 100\%$ CI, 14 df
$t_{\frac{\alpha}{2}} = t_{\frac{0.01}{2}} = t_{0.005} \approx$ (i) **1.76** (ii) **2.98**.

```
qt(0.995,14) # critical value t, 14 df, for 99% CI
```

```
[1] 2.976843
```

iv. *Using formula.*
Thus, the 99% CI for μ is
$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$
(i) $\mathbf{21.6 \pm 2.15 \times \frac{2.97}{\sqrt{15}}}$ (ii) $\mathbf{21.6 \pm 2.15 \times \frac{3.97}{\sqrt{15}}}$ (iii) $\mathbf{21.6 \pm 2.98 \times \frac{2.97}{\sqrt{15}}}$.
which equals
(i) $\mathbf{21.6 \pm 1.29}$ (ii) $\mathbf{21.6 \pm 2.29}$ (iii) $\mathbf{21.6 \pm 3.29} \approx (19.32, 23.88)$.

(d) *Some comments*

i. (i) **True** (ii) **False**. Long 99% CI better than shorter 95% CI in the sense we are more confident 99% contains or "captures" unknown parameter μ. However, 95% CI better than longer 99% CI in the sense, if unknown parameter μ is 95% interval estimate, we are more certain of location of this unknown parameter.

ii. Since sample size is small, we **can** (ii) **cannot** use central limit theorem.

iii. Match columns.

terms	corn example
(a) population	**(a)** average length of 15 plants, $\bar{X}$
(b) sample	**(b)** average length of all plants, μ
(c) statistic	**(c)** lengths of all plants
(d) parameter	**(d)** observed lengths of 15 plants

terms	(a)	(b)	(c)	(d)
corn example				

3. *Population, sample, statistic and parameter: CI for average corn cob length.* Simple random sample of 15 corn cobs is taken. Assume sample SD in length is $s = 2.97$ and, although we typically don't know it, *population* (not *sample*) length is $\mu = 22$ inches. Assume normality.

   (a) *Population* $\mu = 22$ *length*
   Population $\mu = 22$ is a (i) **statistic** (ii) **parameter**.
   Population μ (i) **changes** (ii) **remains same** for every *random* sample.
   Population μ (usually) (i) **known** (ii) **unknown** to us,
   (although we are pretending for this question we do know it.)

   (b) *Sample* $\bar{x}$ *length*
   Sample $\bar{x}$ is a (i) **statistic** (ii) **parameter**.
   Sample $\bar{x}$ (i) **changes** (ii) **remains same** for every *random* sample.
   Sample $\bar{x}$ (usually)(i) **known** (ii) **unknown** to us:
   it may be $\bar{x} = 21.6$ for one sample, but $\bar{x} = 29.8$ for another sample.

   (c) A 95% CI for μ, if $\bar{x} = 21.6$, is
   $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 21.6 \pm 1.96 \frac{2.97}{\sqrt{15}} =$
   (i) $\mathbf{(19.95, 23.24)}$ (ii) $\mathbf{(23.45, 27.80)}$ (iii) $\mathbf{(28.16, 31.44)}$.

```
mean1.t.interval(21.6,2.97,14,0.95) # m: mean, s: SD, n: sample size, 95% t-interval

       Mean  Critical Value Margin of Error       CI lower       CI upper
  21.600000        2.160369        1.714827      19.885173      23.314827
```

   This 95% CI (i) **contains** (ii) **does not contain** $\mu = 22$.

   (d) A 95% CI for μ, if $\bar{x} = 29.8$, is
   $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 29.8 \pm 1.96 \frac{2.97}{\sqrt{15}} =$
   (i) $\mathbf{(19.60, 23.60)}$ (ii) $\mathbf{(23.45, 27.80)}$ (iii) $\mathbf{(28.16, 31.44)}$.

```
mean1.t.interval(29.8,2.97,14,0.95) # m: mean, s: SD, n: sample size, 95% t-interval

       Mean  Critical Value Margin of Error       CI lower       CI upper
  29.800000        2.160369        1.714827      28.085173      31.514827
```

   This 95% CI (i) **contains** (ii) **does not contain** $\mu = 22$.

   (e) If sample average length, $\bar{x}$, changes, corresponding 95% CI,
   $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, (i) **changes** (ii) **remains the same**. More than this,

      i. *all* possible 95% CIs contain $\mu = 22$.
      ii. *none* of all possible 95% CIs contain $\mu = 22$.
      iii. ninety–nine percent of all possible 95% CIs contain $\mu = 22$, and so one percent of all possible 95% CIs do not contain $\mu = 22$.
      iv. ninety–five percent of all possible 95% CIs contain $\mu = 22$, and so five percent of all possible 95% CIs do not contain $\mu = 22$.

   This is demonstrated in figure below.

   (f) Choose true or false.

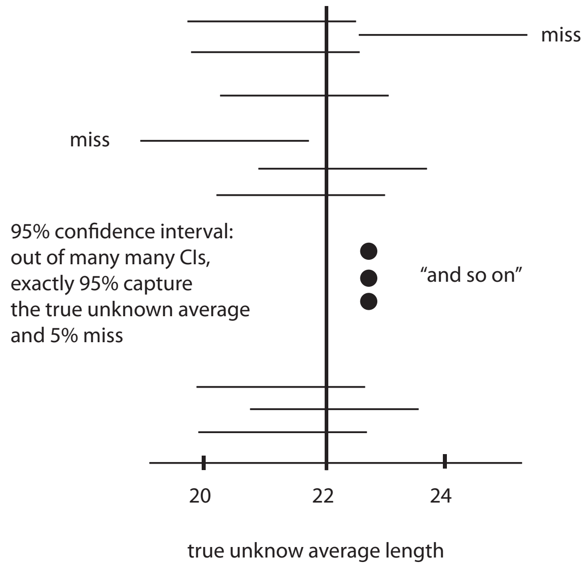

Figure 4.6: Interpreting confidence intervals

(i) **True** (ii) **False**. 95% *chance* (19.95, 23.24) contains μ.
(i) **True** (ii) **False**. 95% *chance* (19.95, 23.24) contains $\bar{x} = 21.6$.
(i) **True** (ii) **False**. 95% *confident* (19.95, 23.24) contains μ.
(i) **True** (ii) **False**. 95% *confident* (19.95, 23.24) contains $\bar{x} = 21.6$.

4.7 Confidence Intervals for a Variance

Let s^2 be the variance of a random sample of size n taken from a normally distributed population with unknown variance σ^2 and $0 < \alpha < 1$. The $(1-\alpha) \cdot 100\%$ confidence interval (CI) for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$$

Also, *one-sided confidence interval estimates* for σ^2 include lower and upper bound respectively:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha}}, \infty \right), \quad \left(0, \frac{(n-1)s^2}{\chi^2_{1-\alpha}} \right).$$

Exercise 4.7 (Confidence Intervals for a Variance)

1. *Estimation for variance: car door and jamb.*
   In a simple random sample of 28 cars, variance in gap between door and jamb is $s^2 = 0.7$ mm^2. Calculate 95% CI. Assume normality with no outliers.

(a) *Using R.* The 95% CI for σ^2 is
(i) $\mathbf{(0.39, 1.22)}$ (ii) $\mathbf{(0.41, 1.25)}$ (iii) $\mathbf{(0.44, 1.30)}$.

```
var1.chi2.interval = function(v,n,conf.level) {
  df = n - 1
  chilower = qchisq((1 - conf.level)/2, df)
  chiupper = qchisq((1 - conf.level)/2, df, lower.tail = FALSE)
  ci.lower <- df * v/chiupper
  ci.upper <- df * v/chilower
  margin.error <- (ci.upper - ci.lower)/2
  dat <- c(v, chilower, chiupper, margin.error, ci.lower, ci.upper)
  names(dat) <- c("Variance", "Lower Crit Val", "Upper Crit Val", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
var1.chi2.interval(0.7,28,0.95) # 95% CI for variance, n = 28
```

```
       Variance  Lower Crit Val  Upper Crit Val Margin of Error        CI lower        CI upper
      0.7000000      14.5733827      43.1945110       0.4296647       0.4375556       1.2968849
```

(b) *Upper* critical value for 95% = $(1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI is
$\chi^2_{\frac{\alpha}{2}} = \chi^2_{\frac{0.05}{2}} = \chi^2_{0.025} =$ (i) **8.7** (ii) **40.1** (iii) **43.2**

```
qchisq(0.975, 27) # 95% upper critical chi-square value
```

```
[1] 43.19451
```

(c) *Lower* critical value for 95% = $(1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI is
$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{1-\frac{0.05}{2}} = \chi^2_{0.975} =$ (i) **14.6** (ii) **40.1** (iii) **43.2**

```
qchisq(0.025, 27) # 95% lower critical chi-square value
```

```
[1] 14.57338
```

(d) Using Table C.4, lower critical value for 95% CI is
$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{1-\frac{0.05}{2}} = \chi^2_{0.975} =$
(i) **between 13.12 and 16.79** (ii) **40.1** (iii) **43.2**

(e) So, 95% CI for variance σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = \left(\frac{(28-1)0.7}{43.2}, \frac{(28-1)0.7}{14.6}\right) =$$

(i) $\mathbf{(0.61, 1.65)}$ (ii) $\mathbf{(0.59, 1.29)}$ (iii) $\mathbf{(0.43, 1.29)}$.

(f) Since 95% CI (0.43, 1.29) does *not* include 0.40, this *indicates* variance in distance between door and jamb (i) **is** (ii) **is not** 0.4 mm^2.

(g) *Population, parameter, sample and statistic.* Match columns.

terms	jamb example
(a) population	**(a)** variance in jamb–door distance, of 28 cars, s^2
(b) sample	**(b)** variance in jamb–door distance, of all cars, σ^2
(c) statistic	**(c)** jamb–door distances, of all cars
(d) parameter	**(d)** jamb–door distances, of 28 cars

terms	(a)	(b)	(c)	(d)
jamb example				

2. *Estimation for variance: machine parts.*
   In a simple random sample of 18 machine parts, variance in lengths is $s^2 = 12^2$. Calculate 90% CI. Assume normality with no outliers.

   (a) *Using R.* The 90% CI for σ^2 is
   (i) $\mathbf{(88.1, 281.3)}$ (ii) $\mathbf{(88.7, 282.3)}$ (iii) $\mathbf{(88.2, 282.3)}$.

   ```
   var1.chi2.interval(12^2,18,0.90) # 90% CI for variance, n = 18

        Variance  Lower Crit Val  Upper Crit Val Margin of Error        CI lower        CI upper
       144.00000         8.67176        27.58711        96.77927        88.73709       282.29563
   ```

   (b) *Upper* critical value for $90\% = (1-\alpha)\cdot 100\% = (1-0.10)\cdot 100\%$ CI is
   $\chi^2_{\frac{\alpha}{2}} = \chi^2_{\frac{0.10}{2}} = \chi^2_{0.05} =$ (i) **8.7** (ii) **27.6** (iii) **43.2**

   ```
   qchisq(0.95, 17) # 90% upper critical chi-square value

   [1] 27.58711
   ```

   (c) *Lower* critical value for $90\% = (1-\alpha)\cdot 100\% = (1-0.10)\cdot 100\%$ CI is
   $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{1-\frac{0.10}{2}} = \chi^2_{0.95} =$ (i) **8.7** (ii) **40.1** (iii) **43.2**

   ```
   qchisq(0.05, 17) # 90% lower critical chi-square value

   [1] 8.67176
   ```

   (d) So, 90% CI for variance σ^2 is (there may round-off error)
   $\left(\frac{(n-1)s^2}{\chi^2_U}, \frac{(n-1)s^2}{\chi^2_L}\right) = \left(\frac{(18-1)12^2}{27.6}, \frac{(18-1)12^2}{8.7}\right) =$
   (i) $\mathbf{(80.5, 101.4)}$ (ii) $\mathbf{(100.5, 104.2)}$ (iii) $\mathbf{(88.7, 281.4)}$.

   (e) Since 90% CI (88.7, 281.4) includes test statistic $13^2 = 169$, this *indicates* variance in lengths (i) **is** (ii) **is not** $\sigma^2 = 13^2$ mm^2.

   (f) Also, 90% CI for *standard deviation* σ is
   $\left(\sqrt{\frac{(n-1)s^2}{\chi^2_U}}, \sqrt{\frac{(n-1)s^2}{\chi^2_L}}\right) = \left(\sqrt{\frac{(18-1)12^2}{27.6}}, \sqrt{\frac{(18-1)12^2}{8.7}}\right) =$
   (i) $\mathbf{(9.4, 16.8)}$ (ii) $\mathbf{(10.5, 14.2)}$ (iii) $\mathbf{(88.7, 281.4)}$.

4.8 Confidence Intervals for Differences

Let x_1 and x_2 be number of successes in two independent samples of size n_1 and n_2 (with $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$) taken two populations with proportions p_1 and p_2. The $(1-\alpha)\cdot 100\%$ *2-proportion z-interval* for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where we assume the samples random and there are at least 5 successes and 5 failures in each sample.

Let $\bar{x}_1$ and $\bar{x}_2$ be the means of two independent samples of size n_1 and n_2 from two populations and means μ_1 and μ_2. The $(1-\alpha)\cdot 100\%$ *2-sample z-interval* for $\mu_1 - \mu_2$, with *known* variances σ_1^2 and σ_2^2, is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

or, with *un*known variances but where it is assumed $\sigma_1^2 = \sigma_2^2$,

$$(\bar{x}_1 - \bar{x}_2) \pm s_p \cdot t_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where the *pooled standard deviation* estimate is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}},$$

and $t_{\frac{\alpha}{2}}$ has $n_1 + n_2 - 2$ degrees of freedom or, with *un*known variances but where it is assumed $\sigma_1^2 \neq \sigma_2^2$,

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and $t_{\frac{\alpha}{2}}$ has the following r degrees of freedom (round down),

$$r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

where either the underlying distribution of both samples are normal with no outliers or if both random sample sizes large ($n_1 \geq 30, n_2 \geq 30$). Also, if the two samples are *dependent* or *paired*, the confidence interval for the difference in two means μ_d is

$$\bar{d} \pm t_{\frac{\alpha}{2}} \left(\frac{s_d}{\sqrt{n}}\right),$$

where either the underlying distribution of *differences* is normal with no outliers or the random sample size is large ($n \geq 30$).

Exercise 4.8 (Confidence Intervals for a Differences)

1. *Inference $p_1 - p_2$, large independent samples: doctors.*
   Calculate 95% 2-proportion z-interval of difference in proportions of male doctors in military and civilian hospitals.

	military (1)	civilian (2)
male doctors	358	6786
total doctors	407	7363

From above, $\hat{p}_1 = \frac{358}{407}$, $\hat{p}_2 = \frac{6786}{7363}$; also
critical value for 95% $= (1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI,
of $z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} \approx$ (i) **1.65** (ii) **1.96** (iii) **2.09**,

```
qnorm(0.975) # critical value z, for 95% CI
```

```
[1] 1.959964
```

and so 95% CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \left(\frac{358}{407} - \frac{6786}{7363}\right) \pm 1.96 \cdot \sqrt{\frac{\frac{358}{407}\left(1-\frac{358}{407}\right)}{407} + \frac{\frac{6786}{7363}\left(1-\frac{6786}{7363}\right)}{7363}} \approx$$

(i) $\mathbf{(-0.054, -0.008)}$ (ii) $\mathbf{(-0.064, -0.009)}$ (iii) $\mathbf{(-0.074, -0.010)}$

```
prop2.interval <- function(x, n, conf.level) {
  x1 <- x[1]; x2 <- x[2]; n1 <- n[1]; n2 <- n[2]
  p.hat1 <- x1/n1; p.hat2 <- x2/n2
  z.crit <- -1*qnorm((1-conf.level)/2)
  margin.error <- z.crit*sqrt(p.hat1*(1-p.hat1)/n1+p.hat2*(1-p.hat2)/n2)
  ci.lower <- p.hat1-p.hat2 - margin.error
  ci.upper <- p.hat1-p.hat2 + margin.error
  dat <- c(p.hat1, p.hat2, z.crit, margin.error, ci.lower, ci.upper)
  names(dat) <- c("p.hat1", "p.hat2", "z crit", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
prop2.interval(c(358,6786), c(407,7363), 0.95) # approx 2-proportion z-test for p, two-sided
```

```
         p.hat1          p.hat2          z crit Margin of Error        CI lower        CI upper
    0.879606880     0.921635203     1.959963985     0.032205624    -0.074233948    -0.009822699
```

Since confidence interval does *not* include (is, in fact, smaller than) zero, this indicates population proportion of male military doctors
(i) **is less than** (ii) **equals** (iii) **is greater than** (iv) **is different from**
the population proportion of male civilian doctors.

2. *CI for $\mu_1 - \mu_2$, independent samples, unknown $\sigma_1^2 = \sigma_2^2$: progesterone.*
A study is conducted to determine cellular response to progesterone in females. Blood cells from four females are injected with progesterone; blood cells from four *different* females are, for comparison purposes, left untreated. Calculate 95% CI. Assume normality with no outliers.

female	progesterone (1)	female	control (2)
1	5.85	5	5.23
2	2.28	6	1.21
3	1.51	7	1.40
4	2.12	8	1.38

```
progesterone <- c(5.85, 2.28, 1.51, 2.12)
control <- c(5.23, 1.21, 1.40, 1.38)
```

From R, $\bar{x}_1 \approx 2.94$, $s_1 \approx 1.97$, $\bar{x}_2 \approx 2.305$, $s_2 \approx 1.95$,

```
m1 <- mean(progesterone); m1; s1 <- sqrt(var(progesterone)); s1
m2 <- mean(control); m2; s2 <- sqrt(var(control)); s2
```

```
> mean(progesterone); sqrt(var(progesterone))
[1] 2.94
[1] 1.968163
> mean(control); sqrt(var(control))
[1] 2.305
[1] 1.95186
```

so pooled standard deviation is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \approx \sqrt{\frac{(4-1)1.97^2 + (4-1)1.95^2}{2+4-2}} \approx$$

(i) **1.95** (ii) **1.96** (iii) **1.97** (which not surprising since $s_1 \approx 1.97$, $s_2 \approx 1.95$)

```
n1 <- length(progesterone); n2 <- length(control)
s12 <- var(progesterone); s22 <- var(control)
sp <- sqrt(((n1-1)*s12 + (n2-1)*s22)/(n1+n2-2)); sp
```

```
[1] 1.96003
```

and critical value for 95% $= (1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI,
with degrees of freedom $= n_1 + n_2 - 2 = 4 + 4 - 2 =$ (i) **4** (ii) **6** (ii) **8**,
so $t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \approx$ (i) **2.31** (ii) **2.45** (iii) **3.09**,

```
qt(0.975,6) # critical t value, 95% CI, using r df
```

```
[1] 2.446912
```

and so 95% CI for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm s_p \cdot t_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (2.94 - 2.305) \pm 1.96 \cdot 2.45 \cdot \sqrt{\frac{1}{4} + \frac{1}{4}} =$$

(i) $\mathbf{(-2.52, 6.49)}$ (ii) $\mathbf{(-2.62, 6.39)}$ (iii) $\mathbf{(-2.76, 4.03)}$

```
mean2.t.interval <- function(m1,m2,s1,s2,n1,n2,conf.level,type) {
  if(type=="same.var") {
    df <- n1+n2-2
    t.crit <- -1*qt((1-conf.level)/2,df)
    sp <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(df))
    margin.error <- sp*t.crit*sqrt(1/n1 + 1/n2)
    ci.lower <- (m1-m2) - margin.error
    ci.upper <- (m1-m2) + margin.error
    dat <- c(m1-m2, df, t.crit, margin.error, ci.lower, ci.upper)
  }
  if(type=="diff.var") {
    df <- ((s1^2/n1 + s2^2/n2)^2)/((1/(n1-1))*(s1^2/n1)^2 + (1/(n2-1))*(s2^2/n2)^2)
    t.crit <- -1*qt((1-conf.level)/2,df)
    margin.error <- t.crit*sqrt(s1^2/n1 + s2^2/n2)
    ci.lower <- (m1-m2) - margin.error
    ci.upper <- (m1-m2) + margin.error
    dat <- c(m1-m2, df, t.crit, margin.error, ci.lower, ci.upper)
  }
  names(dat) <- c("Mean Difference", "df", "Critical Value", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
# approximate 2-mean 95% CI for mu_1 - mu_2, same variance
mean2.t.interval(m1,m2,s1,s2,n1,n2, 0.95,"same.var")
```

```
Mean Difference              df  Critical Value Margin of Error        CI lower        CI upper
       0.635000        6.000000        2.446912        3.391298       -2.756298        4.026298
```

So, since confidence interval *does* include zero, this indicates progesterone population mean cellular response
(i) **is less than** (ii) **equals** (ii) **is greater than** (ii) **is different from**
control population mean cellular response.

3. *Inference for $\mu_1 - \mu_2$, independent samples, unknown $\sigma_1^2 \neq \sigma_2^2$: progesterone.* Same question as before but assume unknown $\sigma_1^2 \neq \sigma_2^2$. Calculate 95% CI.

female	progesterone (1)	female	control (2)
1	5.85	5	5.23
2	2.28	6	1.21
3	1.51	7	1.40
4	2.12	8	1.38

```
progesterone <- c(5.85, 2.28, 1.51, 2.12)
control <- c(5.23, 1.21, 1.40, 1.38)
```

From R, $\bar{x}_1 \approx 2.94$, $s_1 \approx 1.97$, $\bar{x}_2 \approx 2.305$, $s_2 \approx 1.95$ (notice $s_1 \approx s_2$)

```
m1 <- mean(progesterone); m1; s1 <- sqrt(var(progesterone)); s1
m2 <- mean(control); m2; s2 <- sqrt(var(control)); s2
```

```
> mean(progesterone); sqrt(var(progesterone))
[1] 2.94
[1] 1.968163
> mean(control); sqrt(var(control))
[1] 2.305
[1] 1.95186
```

and critical value for 95% $= (1-\alpha)\cdot 100\% = (1-0.05)\cdot 100\%$ CI, with degrees of freedom =

$$r = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{1.97^2}{4}+\frac{1.95^2}{4}\right)^2}{\frac{1}{4-1}\left(\frac{1.97^2}{4}\right)^2+\frac{1}{4-1}\left(\frac{1.95^2}{4}\right)^2} \approx$$

(i) **4** (ii) **6** (ii) **8** (same as when $\sigma_1^2 = \sigma_2^2$)

```
df = 5.9996
```

so $t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \approx$ (i) **2.31** (ii) **2.45** (iii) **3.09**,

```
qt(0.975,6) # critical t value, 95% CI, n1 + n2 - 2 = 6 df

[1] 2.446912
```

and so 95% CI for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = (2.94-2.305) \pm 2.45\cdot\sqrt{\frac{1.97^2}{4}+\frac{1.95^2}{4}} =$$

(i) $\mathbf{(-2.52, 6.49)}$ (ii) $\mathbf{(-2.62, 6.39)}$ (iii) $\mathbf{(-2.76, 4.03)}$

```
mean2.t.interval(m1,m2,s1,s2,n1,n2, 0.95,"diff.var")

Mean Difference              df   Critical Value Margin of Error        CI lower         CI upper
       0.635000        5.999585         2.446953        3.391355       -2.756355         4.026355
```

Since confidence interval *does* include zero, this indicates progesterone population mean cellular response
(i) **is less than** (ii) **equals** (ii) **is greater than** (ii) **is different from**
control population mean cellular response.

4. *Inference for difference in dependent means, μ_d: milk yield.*
A study is conducted to determine effect of "gentech" animal feed on milk yield of 9 cows. Cow 1 is fed a control feed for three months and then gentech feed for next three months for comparison purposes. Other cows are treated in same way. Calculate 95% CI of mean *paired* differences in milk yield. Fill in blanks.

cow	gentech (1)	control (2)	differences, d_i
1	62	54	______
2	45	43	______
3	53	55	______
4	35	39	______
5	71	65	______
6	64	62	______
7	63	56	______
8	57	50	______
9	43	52	______

```
gentech <- c(62, 45, 53, 35, 71, 64, 63, 57, 43)
control <- c(54, 43, 55, 39, 65, 62, 56, 50, 52)
diff <- gentech - control; diff
```

```
[1]  8  2 -2 -4  6  2  7  7 -9
```

$\bar{d} \approx$ (i) **1.41** (ii) **1.89** (iii) **2.52**,
$s_d \approx$ (i) **5.47** (ii) **5.86** (iii) **6.52**,

```
mean(diff); sqrt(var(diff))
```

```
[1] 1.888889
[1] 5.861835
```

with $n - 1 = 9 - 1 =$ (i) **6** (ii) **7** (ii) **8** degrees of freedom,
and critical value $95\% = (1 - \alpha) \cdot 100\% = (1 - 0.05) \cdot 100\%$ CI,
so $t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \approx$ (i) **2.31** (ii) **2.53** (iii) **3.09**,

```
qt(0.975,8) # critical t value, 95% CI, nd - 1 = 9 - 1 = 8 df
```

```
[1] 2.306004
```

and so 95% CI for μ_d is

$$\bar{d} \pm t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} = 1.89 \pm 2.31 \times \frac{5.86}{\sqrt{8}} =$$

(i) $\mathbf{(-2.52, 6.49)}$ (ii) $\mathbf{(-2.62, 6.39)}$ (iii) $\mathbf{(-2.72, 6.29)}$

```
mean1.t.interval <- function(m,s,n,conf.level)
{
  t.crit <- -1*qt((1-conf.level)/2,n-1)
  margin.error <- t.crit*s/sqrt(n)
  ci.lower <- m - margin.error
  ci.upper <- m + margin.error
  dat <- c(mean, t.crit, margin.error, ci.lower, ci.upper)
  names(dat) <- c("Mean", "Critical Value", "Margin of Error", "CI lower", "CI upper")
  return(dat)
}
mean1.t.interval(1.889,5.8618,9,0.95) # m: mean, s: SD, n: sample size, 95% t-interval
```

```
          Mean  Critical Value Margin of Error        CI lower        CI upper
      1.889000        2.306004        4.505778       -2.616778        6.394778
```

Since confidence interval *does* include zero, this indicates
gentech population mean milk yield
(i) **is less than** (ii) **equals** (iii) **is greater than** (iv) **is different from**
control population mean milk yield.

4.9 Sample Size

The length of a confidence interval (equivalently, margin of error) can be controlled by sample size, in particular, the larger the sample size, the smaller, more accurate, the confidence interval.

The sample size necessary to achieve a required margin of error, E, with a given level of confidence in a confidence interval of proportion p is determined using formula, if *prior* $\hat{p}$ available,

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2,$$

and if prior $\hat{p}$ unavailable,

$$n = \frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2.$$

The sample size necessary to achieve a required margin of error, E, with a given level of confidence in a confidence interval of mean μ is determined using formula

$$n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{E}\right)^2.$$

where, if σ^2 is unknown, using approximation

$$\sigma \approx \frac{\max - \min}{4}$$

Exercise 4.9 (Sample Size)

1. *Sample size for proportion p: credit card purchase slips.*

   (a) *With prior $\hat{p}$: purchase slips.*
   In an initial simple random sample, twenty–five (25) of 100 purchase slips chosen are Visa. What is sample size, n, required to estimate proportion Visa purchase slips, p, to within margin of error of $E = 0.01$ with 85% confidence? Here

   $$n = \hat{p}(1-\hat{p})\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 = \left(\frac{25}{100}\right)\left(\frac{75}{100}\right)\left(\frac{1.44}{0.01}\right)^2 \approx$$

   (i) **3886** (ii) **5184** (ii) **5470**.

```
n.prop <- function(p.hat,margin.error,conf.level)
{
  z.crit <- -1*qnorm((1-conf.level)/2)
  p.hat*(1-p.hat)*z.crit^2/margin.error^2
}
n.prop(0.25,0.01,0.85) # n for prior p-hat = 0.25, E = 0.01, 85% confidence
```

```
[1] 3885.47
```

(b) *Sample size for proportion p without prior $\hat{p}$: purchase slips.*
What is sample size, n, required to estimate proportion Visa purchase slips, p, to within margin of error of $E = 0.01$ with 85% confidence? Here

$$n = \frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 = \frac{1}{4}\left(\frac{1.44}{0.01}\right)^2 \approx$$

(i) **4409** (ii) **5181** (iii) **5470**.

```
n.prop(0.5,0.01,0.85) # n required no prior (max = 0.5), E = 0.01, 85% confidence
```

```
[1] 5180.627
```

Without prior $\hat{p} = 0.25$, sample size
(i) **decreases** (ii) **remains same** (iii) **increases**
from $n \approx 3886$ to $n \approx 5181$.

2. *Sample size for mean μ: corn cob lengths.*

(a) What sample size, n, required to estimate average corn cob length , μ, to within margin of error $E = 0.08$ with 95% confidence? Assume $\sigma = 0.25$.

$$n = \left(\frac{z_{\frac{\alpha}{2}} s}{E}\right)^2 = \left(\frac{z_{0.025}\sigma}{E}\right)^2 = \left(\frac{1.96 \cdot 0.25}{0.08}\right)^2 \approx$$

(i) **37** (ii) **38** (iii) **39**.

```
n.mean <- function(s,margin.error,conf.level)
{
  z.crit <- -1*qnorm((1-conf.level)/2)
  s^2*z.crit^2/margin.error^2
}
n.mean(0.25,0.08,0.95)
```

```
[1] 37.51425
```

(b) *Increase margin of error, E.*
What sample size, n, required to estimate average corn cob length , μ, to within margin of error $E = 0.16$ with 95% confidence? Assume $\sigma = 0.25$.

$$n = \left(\frac{z_{\frac{\alpha}{2}} s}{E}\right)^2 = \left(\frac{z_{0.025}\sigma}{E}\right)^2 = \left(\frac{1.96 \cdot 0.25}{0.16}\right)^2 \approx$$

(i) **9** (ii) **10** (iii) **11**.

```
n.mean(0.25,0.16,0.95) # n for sigma = 0.25, E = 0.16, 95% confidence
```

```
[1] 9.378562
```

When margin of error doubled, from $E = 0.08$ to $E = 0.16$, sample size
(i) **quartered** (ii) **halved** (iii) **doubled** from $n = 38$ to $n = 10$.
Less data gives less accurate, wider, CI.

4.10 Assessing Normality

Exercise 4.10 (Assessing Normality)
Check assumptions for average length of corn cobs, raw data.
Corn cob lengths for $n = 15 < 30$ cobs, chosen at random, are noted.

$$18, 23, 24, 20, 21, 19, 27, 24, 19, 20, 25, 20, 18, 26, 20$$

```
x <- c(18, 23, 24, 20, 21, 19, 27, 24, 19, 20, 25, 20, 18, 26, 20)
```

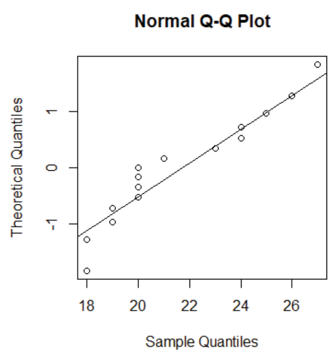

Figure 4.7: Normal probability plot for cob lengths

Normal probability plot indicates cob lengths (i) **normal** (ii) **not normal** because scatter plot (more or less) linear.

```
x <- c(18, 23, 24, 20, 21, 19, 27, 24, 19, 20, 25, 20, 18, 26, 20)
qqnorm(x,datax=TRUE) # QQ plot
qqline(x,datax=TRUE) # QQ line
```

Since classifying normality of data using a normal probability plot is somewhat inexact, we will later look at a measure of linearity called *correlation*, ρ, a number between -1 (perfectly negatively linear) and +1 (perfectly positively linear), which will provide further information to help us decide if the normal probability plot is linear or not.