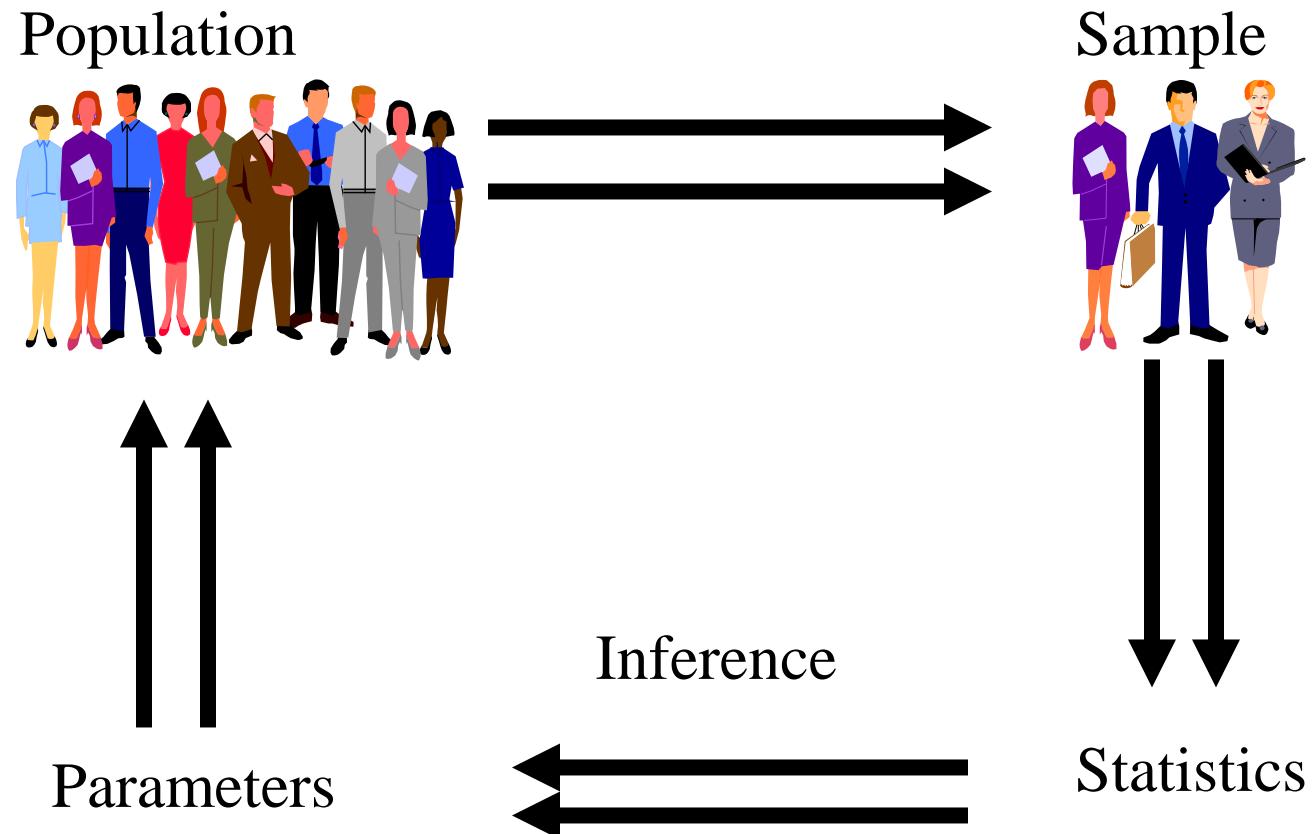


Chapter 6 – Point estimation

Outline

- ① General concepts
- ② Methods of Point Estimation



General concepts of point estimation

- A point estimate of a parameter θ (ex, μ, σ^2, \dots) is a single number : written $\hat{\theta}$.
- The corresponding random variable is called the estimator : written as $\hat{\Theta}$.
- For many parameters there is a simple and obvious estimator.
 - For binomial data with parameter p : $\hat{P} = X/n$, the proportion of successes.
 - For a sample from a distribution with mean μ and variance σ^2 :
$$\hat{\mu} = \bar{X} = \frac{X_1 + \dots + X_n}{n} \text{ with value } \bar{x}.$$

$$\hat{\sigma}^2 = S^2 \text{ with value } s^2.$$
- There can be several possible estimators for parameters such as μ and σ^2 .

Example 6.1

An automobile manufacturer has developed a new type of bumper.

The manufacturer has used this bumper in 25 controlled crashes against a wall.

The parameter to be estimated is

p = the proportion of all such crashes that result in no damage.

Let

X = the number of crashes that result in no visible damage to the automobile.

If X is observed to be $x=15$, then

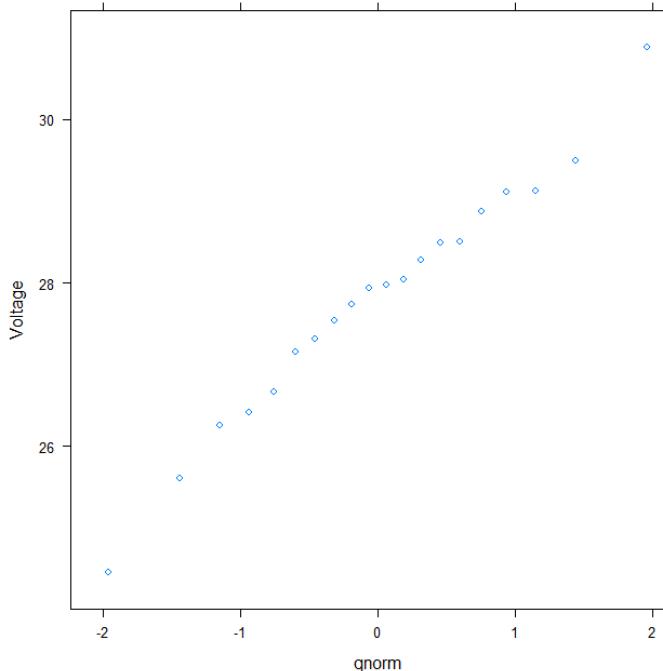
$$\text{estimator } \hat{P} = \frac{x}{n}, \quad \text{estimate } \hat{p} = \frac{x}{n} = \frac{15}{25} = 0.6$$

Example 6.2 : Dielectric breakdown voltage for epoxy resin

A normal probability plot shows that the data closely follow a normal shape

```
> str(xmp06.02)
```

```
> qqmath(~Voltage, xmp06.02)
```



```
> sort(xmp06.02$Voltage)
```

```
[1] 24.46 25.61 26.25 26.42 26.66 27.15 27.31 27.54 27.74 27.94 27.98 28.04
```

```
[13] 28.28 28.49 28.50 28.87 29.11 29.13 29.50 30.88
```

Example 6.2 (cont'd)

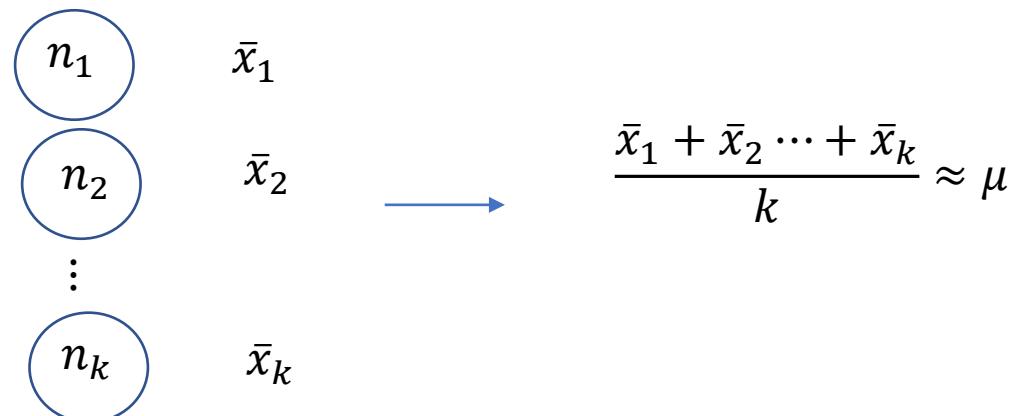
- We could choose several ways of estimating the population mean, μ ,
 - The sample mean, \bar{X} : $\bar{x} = \frac{\sum x_i}{n} = \frac{555.86}{20} = 27.793$
 - The sample median, \tilde{X} : $\tilde{x} = \frac{27.94+27.98}{2} = 27.96$
 - The average of the largest and smallest values :
$$[\min(x_i) + \max(x_i)]/2 = (24.46 + 30.88)/2 = 27.67.$$
 - A 10% trimmed mean : $\bar{x}_{tr(10)} = \frac{555.86 - 24.46 - 25.61 - 29.50 - 30.88}{16} = 27.838$
- It is helpful to have some criteria for comparing different estimators.
- The expected values and variances of different estimators can form a basis for comparisons.

Unbiasedness

- An estimator $\hat{\theta}$ for the parameter θ is said to be unbiased if

$$E[\hat{\theta}] = \theta$$

- The sample mean \bar{X} is an unbiased estimator for μ if we can assume that X_1, \dots, X_n are a random sample (independent and identically distributed)
- The sample median \tilde{X} is an unbiased estimator for μ if we can assume that X_1, \dots, X_n are a random sample and the distribution of the X_i s is continuous and symmetric.



- In Example 6.1, the sample proportion X/n was used as an estimator of p where X had a binomial distribution with parameters n and p .

Thus

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

- When X is a binomial random variable with parameters n and p , the sample proportion \hat{p} is an unbiased estimator of p

Unbiasedness

- The sample variance $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ is an unbiased estimator for σ^2 , $\sqrt{S^2}$ is not an unbiased estimator for σ .

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2)}{n-1} = \frac{1}{n-1} [\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2] \quad \left(\bar{X} = \frac{\sum X_i}{n}\right) \\ &= \frac{1}{n-1} [\sum X_i^2 - 2\frac{(\sum X_i)^2}{n} + \frac{(\sum X_i)^2}{n}] = \frac{1}{n-1} [\sum X_i^2 - \frac{(\sum X_i)^2}{n}] \\ E(S^2) &= \frac{1}{n-1} \left\{ \sum E(X_i^2) - \frac{1}{n} E[(\sum X_i)^2] \right\} \\ &= \frac{1}{n-1} \left\{ \sum (\sigma^2 + \mu^2) - \frac{1}{n} \{V(\sum X_i) + [E(\sum X_i)]^2\} \right\} \quad (Y = \sum X_i, \quad E(Y^2) = V(Y) + E(Y)^2) \\ &= \frac{1}{n-1} \left\{ n\sigma^2 + n\mu^2 - \frac{1}{n} n\sigma^2 - \frac{1}{n} (n\mu)^2 \right\} \\ &= \frac{1}{n-1} \{n\sigma^2 - \sigma^2\} = \sigma^2 \end{aligned}$$

Variances of estimators

- Given two unbiased estimators, we generally prefer the one with the smaller variance.
- Occasionally it is possible to prove mathematically that an estimator is a minimum variance unbiased estimator (MVUE).

This means it has the minimum variance among the class of unbiased estimators.

- If X_1, \dots, X_n is a random sample from a normal distribution with mean μ and variance σ^2 , then \bar{X} is an MVUE for μ .
- This is about the only practical case where you can establish this property.
- In general the desirability of an estimator depends on the form of the underlying distribution. When working with real data we don't know what the distribution is.
- We can compare estimators on the basis of simulation.

Comparing estimators for the mean μ , normal distribution $N(6, 1.2^2)$

```
> k <- 50000
> n <- 25
> mns <- meds <- extravg <- trmns <- numeric(k)
> for (i in 1:k) {
+   samp <- rnorm(n, mean=6, sd=1.2)
+   mns[i] <- mean(samp)
+   meds[i] <- median(samp)
+   extravg[i] <- mean(range(samp))
+   trmns[i] <- mean(samp, trim=0.05)
+ }
> resultsN <- list(means = mns, medians = meds, extravg = extravg, trimmed = trmns)
> sapply(resultsN, mean)
  means      medians      extravg      trimmed
6.000328  6.001225  5.998205  6.000513
> sapply(resultsN, sd)
  means      medians      extravg      trimmed
0.2401946 0.2979679 0.4367173 0.2422739
```

Example 6.2 (cont'd)

- We could choose several ways of estimating the population mean, μ ,
 - The sample mean, \bar{X} : $\bar{x} = \frac{\sum x_i}{n} = \frac{555.86}{20} = 27.793$
 - The sample median, \tilde{X} : $\tilde{x} = \frac{27.94+27.98}{2} = 27.96$
 - The average of the largest and smallest values :
$$[\min(x_i) + \max(x_i)]/2 = (24.46 + 30.88)/2 = 27.67.$$
 - A 10% trimmed mean : $\bar{x}_{tr(10)} = \frac{555.86 - 24.46 - 25.61 - 29.51 - 30.88}{16} = 27.838$
- It is helpful to have some criteria for comparing different estimators.
- The expected values and variances of different estimators can form a basis for comparisons.

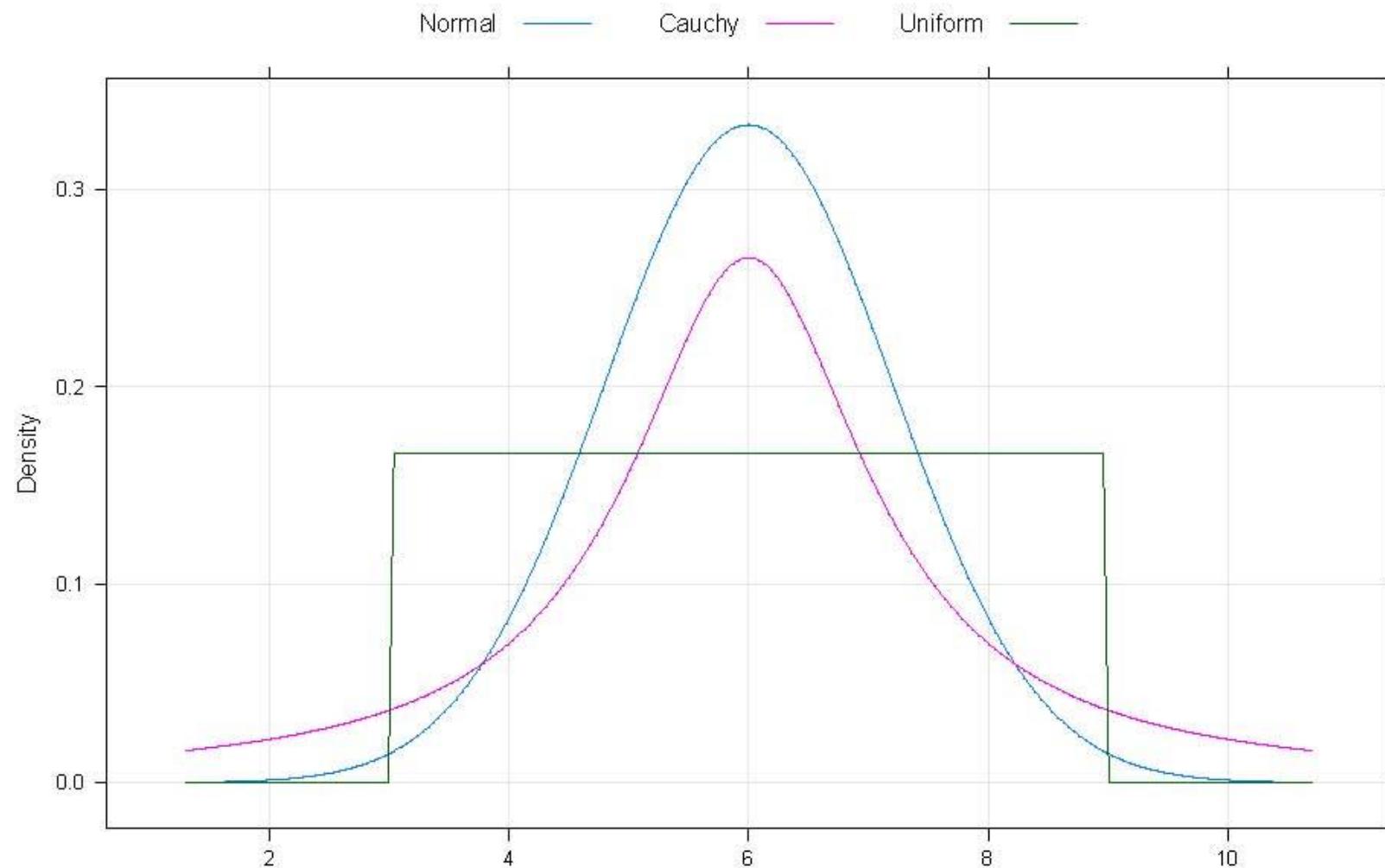
Comparing estimators for the location, Cauchy distribution

```
> for (i in 1:k) {  
+   samp <- rcauchy(n, location=6, scale=1.2)  
+   mns[i] <- mean(samp)  
+   meds[i] <- median(samp)  
+   extravg[i] <- mean(range(samp))  
+   trmns[i] <- mean(samp, trim=0.05)  
+ }  
> resultsN <- list(means = mns, medians = meds, extravg = extravg, trimmed =  
+   trmns)  
> sapply(resultsN, mean)  
    means      medians      extravg      trimmed  
  5.671262  6.001956  1.846844  6.003820  
> sapply(resultsN, sd)  
    means      medians      extravg      trimmed  
  95.96416   0.40234  1198.19459   1.71752
```

Comparing estimators for the mean μ , Uniform distribution

```
> for (i in 1:k) {  
+   samp <- runif(n, min=3, max=9)  
+   mns[i] <- mean(samp)  
+   meds[i] <- median(samp)  
+   extravg[i] <- mean(range(samp))  
+   trmns[i] <- mean(samp, trim=0.05)  
+ }  
> resultsN <- list(means = mns, medians = meds, extravg = extravg, trimmed =  
+   trmns)  
> sapply(resultsN, mean)  
    means     medians     extravg     trimmed  
6.000695 6.001103 6.000357 6.000724  
> sapply(resultsN, sd)  
    means     medians     extravg     trimmed  
0.34762  0.57853  0.16148  0.37156
```

Comparison of the distributions



The standard error of an estimator

- In addition to the point estimate, we should report some measure of the precision of this estimate.
- The standard deviation of the estimator is a reasonable value to report.

Unfortunately, it will usually depend on the values of unknown parameters.

 - For a binomial model, the estimator $\hat{P} = X/n$ of the probability of success, has a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$, which depends on the parameter we are trying to estimate.
 - To estimate μ from a random sample that seems close to a normal distribution, we use the estimator \bar{X} , whose standard deviation, $\frac{\sigma}{\sqrt{n}}$, depends on another unknown parameter σ .
- If we use estimates of the unknown parameters in the formula for the standard deviation we obtain the standard error of the estimator, which is the estimated standard deviation of the estimator.

$$\text{s.e.}(\hat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{s.e.}(\bar{X}) = \frac{s}{\sqrt{n}}$$

Example 6.10

The standard deviation of $\hat{P} = \frac{x}{n}$ is

$$\sigma_{\hat{P}} = \sqrt{V\left(\frac{X}{n}\right)} = \sqrt{\frac{V(X)}{n^2}} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}$$

Since p and $q = 1 - p$ are unknown, we substitute $\hat{p} = \frac{x}{n}$ and $\hat{q} = 1 - \frac{x}{n}$ into $\sigma_{\hat{P}}$, yielding the estimated standard error $\hat{\sigma}_{\hat{P}} = \sqrt{\hat{p}\hat{q}/n} = \sqrt{(0.6)(0.4)/25} = 0.098$

```

> x <- c(132.0, 129.0, 120.0, 113.2, 105.0, 92.0, 84.0, 83.2, 88.4, 59.0, 80.0, 81.5, 71.0, 69.2)
> y <- c(46.0, 48.0, 51.0, 52.1, 54.0, 52.0, 59.0, 58.7, 61.6, 64.0, 61.4, 54.6, 58.8, 58.0)
> plot(x, y)
> f <- lm(y~x)
> summary(f)

```

Call:

`lm(formula = y ~ x)`

Residuals:

Min	1Q	Median	3Q	Max
-3.9488	-1.5665	0.6817	1.0846	4.8974

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	75.21243	2.98363	25.208	9.22e-12 ***
x	-0.20939	0.03109	-6.734	2.09e-05 ***

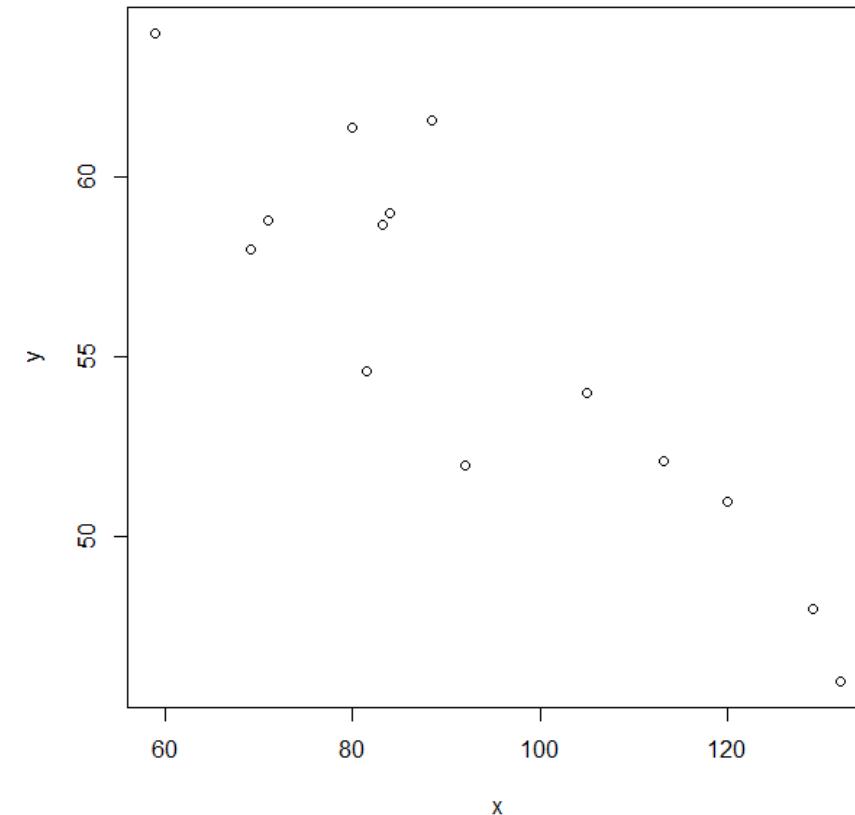
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.564 on 12 degrees of freedom

Multiple R-squared: 0.7908, Adjusted R-squared: 0.7733

F-statistic: 45.35 on 1 and 12 DF, p-value: 2.091e-05

- Regression equation : $Y = 75.21243 - 0.20939X$



Formulating estimators

- We have seen how to compare different estimators for a parameter based on their expected values and variances.
- Knowing how to compare estimators doesn't tell us how to formulate an estimator.
- We need to decide how to find one or more estimators for the parameters in a distribution before we can compare them.

method of moments (MoM)

maximum likelihood estimators (mle's).

- MoM is interesting from the historical perspective but is not in wide use today.
- Generally if we are faced with an unknown estimation situation, we will derive the maximum likelihood estimators for the parameters.

The method of moments

- Early in the history of statistics a lot of attention was paid to the moments of a distribution (mean, variance, skewness, kurtosis)

$$\text{variance} : E[(X - \mu)^2]$$

$$\text{skewness} : E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \quad \text{kurtosis} : E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

- Estimators were formulated by equating the observed moments (i.e., \bar{x}, s^2, \dots) to the theoretical moments, which are functions of the parameters. The number of equations should be equal to the number of parameters.
- This approach is not unreasonable but also it is not clear that it should give a good result.
- Suppose we consider the exponential distribution with rate $\lambda > 0$ (see p.165)

$$f(x ; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then $E[X] = 1/\lambda$ so the MoM estimator of λ is

$$\bar{X} = 1/\lambda \rightarrow \hat{\lambda} = 1/\bar{X} \quad \text{and} \quad \hat{\lambda} = 1/\bar{x}.$$

What could be wrong with the MoM estimators?

- One problem with the MoM estimators is that the math can get pretty ugly.
Suppose you want to estimate the parameters α and β in a Weibull distribution (see pp. 171-173). You start with the equations

$$\mu = \beta \Gamma\left(1 + \frac{1}{\alpha}\right), \sigma^2 = \beta^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right) \right]^2 \right] \rightarrow \bar{x} = \beta \Gamma\left(1 + \frac{1}{\alpha}\right), s^2 = \beta^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right) \right]^2 \right]$$

Solving two equations like that for α and β is not fun

- The values of the estimators don't have to be consistent with the distribution.
- Suppose you have a uniform density on $[0, \theta]$ where θ is the unknown parameter.
 $E(X) = \theta/2$

It's MoM estimator is $\hat{\theta} = 2\bar{X}$. (from $\bar{X} = \theta/2$)

However, it is possible that in a given data set we could have $\hat{\theta} < \max(x_i)$ and we know that θ must be greater than all the $x_i, i = 1, \dots, n$.

Maximum likelihood estimators

- R.A. Fisher suggested that a way to avoid some of the problems with MoM estimators was to consider the joint density of the responses as a function of the parameters with the data fixed. He called this the likelihood of the parameters given the data. That is,

$$L(\theta|x) = f(x|\theta)$$

where L is the likelihood, f is the probability density, x is the vector of responses and θ is the vector of parameters for the distribution.

- This is the opposite of the way we usually interpret this expression (as a function of the data with the parameters fixed).
- Fisher said that we should choose the estimates of the parameters as those values that provide the greatest likelihood of seeing the data we did. That is,

$$\hat{\theta} = \arg \max_{\theta} L(\theta|x)$$

Maximum Likelihood Estimation – Example 6.16

Let X_1, \dots, X_n be a random sample from an exponential distribution with parameter λ . Because of the independence, the likelihood function is a product of the individual pdf's:

$$L(\lambda|x) = f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdots (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$

The natural logarithm of the likelihood function is

$$l(\lambda|x) = \log L(\lambda|x) = n \ln(\lambda) - \lambda \sum x_i$$

Equating $[\frac{d}{d\lambda} l(\lambda|x)]$ to zero results in

$$\frac{n}{\lambda} - \sum x_i = 0, \text{ or } \lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}} \quad \left(\frac{d}{d\lambda} \ln[f(x)] = \frac{f'(x)}{f(x)} \right)$$

Thus the mle is

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

- Example 6.11 gives data on breakdown voltages that are assumed to come from an exponential distribution.

41.53, 18.73, 2.99, 30.34, 12.33, 117.52, 73.02, 223.63, 4.00, 26.78

$$\begin{aligned}
 L(\lambda|x) &= f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdots (\lambda e^{-\lambda x_n}) \\
 &= (\lambda e^{-41.53\lambda}) \cdots (\lambda e^{-26.78\lambda}) = \lambda^{10} e^{-550.97\lambda}
 \end{aligned}$$

Maximum Likelihood Estimation

Definition

Let X_1, \dots, X_n have a joint pmf or pdf

$$f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \quad (1)$$

where the parameters $\theta_1, \dots, \theta_m$ have unknown values. When x_1, \dots, x_n are observed sample values and (1) is regarded as a function of $\theta_1, \dots, \theta_m$, it is called the likelihood function.

$$L(\theta_1, \dots, \theta_m | x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$$

Definition

The maximum likelihood estimates (mle's) $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of the θ_i 's that maximize the likelihood function, so that

$$L(\hat{\theta}_1, \dots, \hat{\theta}_m | x_1, \dots, x_n) \geq L(\theta_1, \dots, \theta_m | x_1, \dots, x_n)$$

Maximum Likelihood Estimation in practice

- In most cases it is easier to optimize the logarithm of the likelihood than it is to optimize the likelihood.

The optimum occurs at the same place (i.e. the estimate $\hat{\theta}$ is the same).

That is,

$$\hat{\theta} = \arg \max_{\theta} L(\theta|x) = \arg \max_{\theta} l(\theta|x)$$

where $l(\theta|x) = \log L(\theta|x)$

- Most optimization software is designed to minimize a function.

Thus we re-express the mle's as

$$\hat{\theta} = \arg \min_{\theta} -l(\theta|x)$$

Maximum likelihood estimator example

- Suppose that we obtained 8 heads among 20 coin tosses.

Calculate the maximum likelihood estimator.

$$L(p; x) = P(X = 8; p) = \binom{20}{8} p^8 (1-p)^{12} \approx p^8 (1-p)^{12}$$

$$l(p; x) \approx \log L(p; x) = 8 \log p + 12 \log(1-p) \quad \left(\frac{d}{d\lambda} \ln[f(x)] = \frac{f'(x)}{f(x)} \right)$$

$$\frac{d}{dp} l(p; x) = \frac{8}{p} + \frac{-12}{1-p} = 0 \rightarrow \frac{8}{p} = \frac{12}{1-p} \rightarrow 8(1-p) = 12p \rightarrow 8 = 20p \rightarrow p = \frac{8}{20}$$

- Suppose that we obtained x heads among n coin tosses.

Then the maximum likelihood estimator of p is $\frac{x}{n}$.

Maximum likelihood estimators

- R.A. Fisher suggested that a way to avoid some of the problems with MoM estimators was to consider the joint density of the responses as a function of the parameters with the data fixed. He called this the likelihood of the parameters given the data. That is,

$$L(\theta|x) = f(x|\theta)$$

where L is the likelihood, f is the probability density, x is the vector of responses and θ is the vector of parameters for the distribution.

- This is the opposite of the way we usually interpret this expression (as a function of the data with the parameters fixed).
- Fisher said that we should choose the estimates of the parameters as those values that provide the greatest likelihood of seeing the data we did. That is,

$$\hat{\theta} = \arg \max_{\theta} L(\theta|x)$$

Maximum Likelihood Estimation – Example 6.16

Let X_1, \dots, X_n be a random sample from an exponential distribution with parameter λ . Because of the independence, the likelihood function is a product of the individual pdf's:

$$L(\lambda|x) = f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdots (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$

The natural logarithm of the likelihood function is

$$l(\lambda|x) = \log L(\lambda|x) = n \ln(\lambda) - \lambda \sum x_i$$

Equating $[\frac{d}{d\lambda} l(\lambda|x)]$ to zero results in

$$\frac{n}{\lambda} - \sum x_i = 0, \text{ or } \lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

Thus the mle is

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

Maximum Likelihood Estimation – Example 6.17

Let X_1, \dots, X_n be a random sample from the normal distribution

$$\begin{aligned} L(\mu, \sigma^2 | x) &= f(x_1, \dots, x_n; \mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_2-\mu)^2/(2\sigma^2)} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)} \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\sum_{i=1}^n (x_i - \mu)^2 / (2\sigma^2)} \end{aligned} \tag{3}$$

The natural logarithm of the likelihood function is

$$l(\mu, \sigma^2 | x) = \log L(\mu, \sigma^2 | x) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

To find the maximizing values of μ and σ^2 , we must take the partial derivatives of $\ln(L)$ with respect to μ and σ^2 , equate them to zero, and solve the resulting two equations.

Maximum Likelihood Estimation – Example 6.17

$$\frac{\partial l(\mu, \sigma^2 | x)}{\partial \mu} = \frac{\partial}{\partial \mu} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] = \frac{-2 \sum_{i=1}^n (x_i - \mu)(-1)}{2\sigma^2} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = 0$$

$$\rightarrow \sum_{i=1}^n (x_i - \mu) = 0 \rightarrow n\mu = \sum_{i=1}^n x_i \quad \left(\frac{d}{dx} (f(x)^a) = af(x)^{a-1} f'(x) \right)$$

$$\rightarrow \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}$$

$$\frac{\partial l(\mu, \sigma^2 | x)}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \left[-\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \quad \frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}$$

$$= -\frac{n}{2} \frac{2\pi}{2\pi\sigma^2} - \sum_{i=1}^n (x_i - \mu)^2 \frac{-2}{(2\sigma^2)^2} \quad \left(\frac{d}{dx} \ln[f(x)] = \frac{f'(x)}{f(x)} \right)$$

$$= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2(\sigma^2)^2} = -n\sigma^2 + \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

Using *R* to get mle's

- For a given set of data we can evaluate the logarithm of the density for a distribution with the d"name" function. We add log=TRUE to the call to get the logarithm of the density, then sum the result and add a negative sign.

```
> dexp(3, 0.5, log=T)      #log(0.5exp(-0.5 × 3))
```

```
> log(dexp(3, 0.5))
```

- Example 6.11 gives data on breakdown voltages that are assumed to come from an exponential distribution. The method of moments estimator for the parameter λ is the reciprocal of the sample mean.

```
> Time <- c(41.53, 18.73, 2.99, 30.34, 12.33, 117.52, 73.02, 223.63, 4.00, 26.78)
```

```
> Time
```

```
[1] 41.53 18.73 2.99 30.34 12.33 117.52 73.02 223.63 4.00 26.78
```

```
> 1/mean(Time)
```

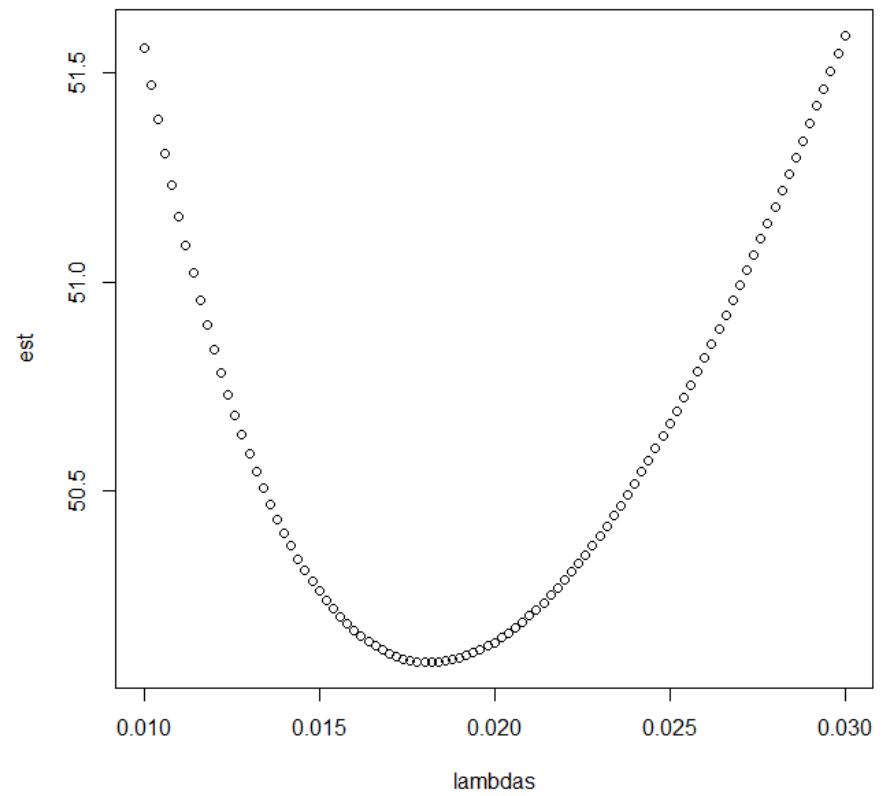
```
[1] 0.0181531
```

Using *R* to get mle's

$$L(\lambda|x) = f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdots (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$

$$-\log(L(\lambda|x)) = -\sum_{i=1}^n \log(\lambda e^{-\lambda x_i})$$

```
> k <- 101  
  
> negloglik <- function(lambda){  
+  -sum(dexp(Time, rate=lambda, log=TRUE))  
+ }  
  
> lambdas <- seq(0.01, 0.03, length=k)  
  
> est <- numeric(k)  
  
> for(i in 1:k){  
+   est[i] = negloglik(lambdas[i])  
+ }  
  
> plot(lambdas, est)
```



Creating the estimates

- *R* provides several optimizer functions that can minimize a function like the negloglik function defined above. One of these, called `nlminb`, allows us to specify bounds on the parameters. Its arguments are the initial guess or starting estimate for the parameter, the function to minimize and, optionally, lower or upper bounds.

```
> str(optres <- nlminb(0.001, negloglik, lower=0))
```

List of 6

```
$ par      : num 0.0182
$ objective : num 50.1
$ convergence: int 0
$ iterations : int 11
$ evaluations: Named int [1:2] 18 14
..- attr(*, "names")= chr [1:2] "function" "gradient"
```

```
$ message   : chr "relative convergence (4)"
```

```
> optres$par
```

```
[1] 0.0181531
```

- Computation of the value of x that yields the minimum of a function.

```
> cubic <- function(x) {  
+   x^2 + 2*x + 1  
+ }
```

```
> nlmnb(0, cubic)
```

```
$par
```

```
[1] -1
```

```
$objective
```

```
[1] 0
```

- Mle for the normal distribution :

```
> x <- rnorm(30, 3, 2)
> negloglik <- function(par) -sum(dnorm(x, mean=par[1], sd=par[2], log=TRUE))
> str(optres <- nlmnb(c(1,1), negloglik, lower=c(-100, 0)))
```

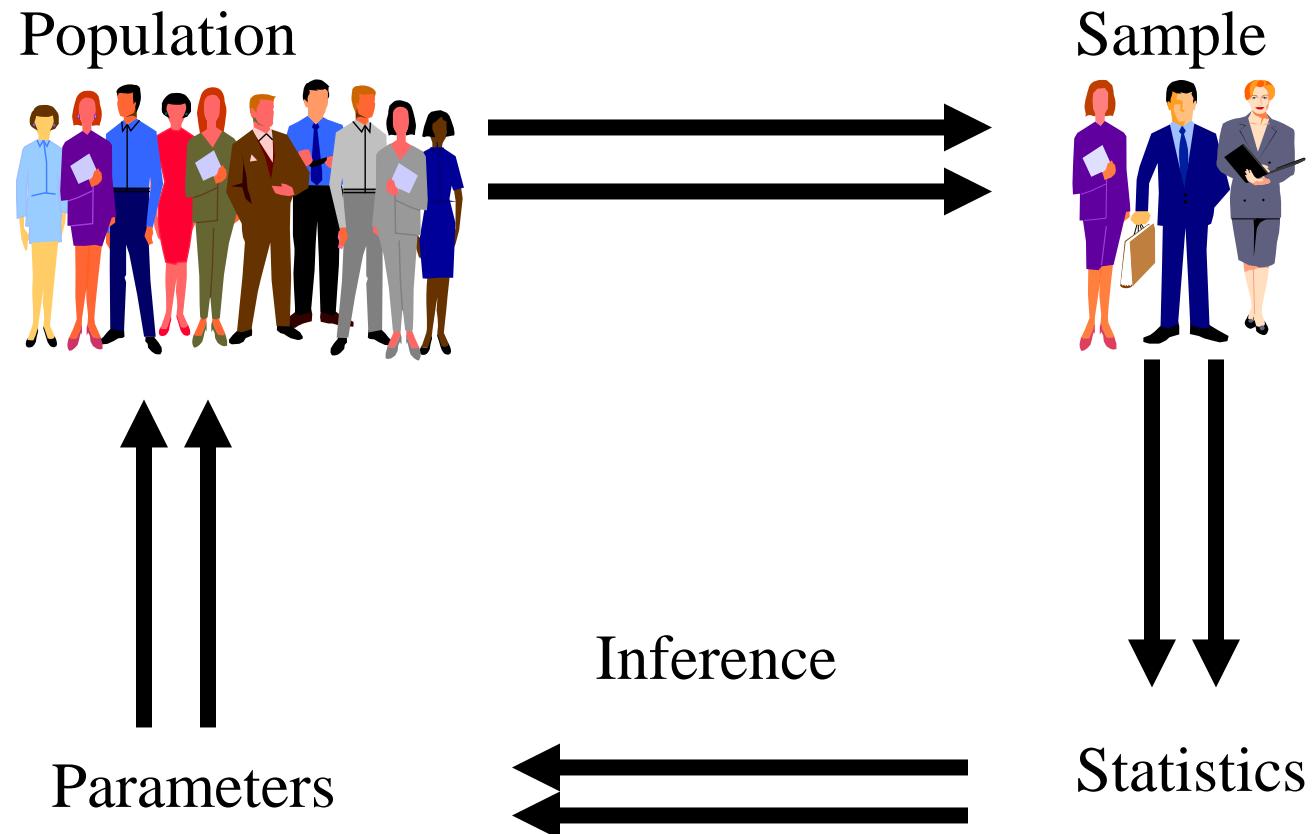
- Mle for the negative binomial distribution :

```
> x <- rnbinom(1000, mu = 10, size = 10)
> hdev <- function(par)
+   -sum(dnbinom(x, mu = par[1], size = par[2], log = TRUE))
> nlmnb(c(9, 12), hdev)
```

Chapter 7 : Statistical Intervals Based on a Single Sample

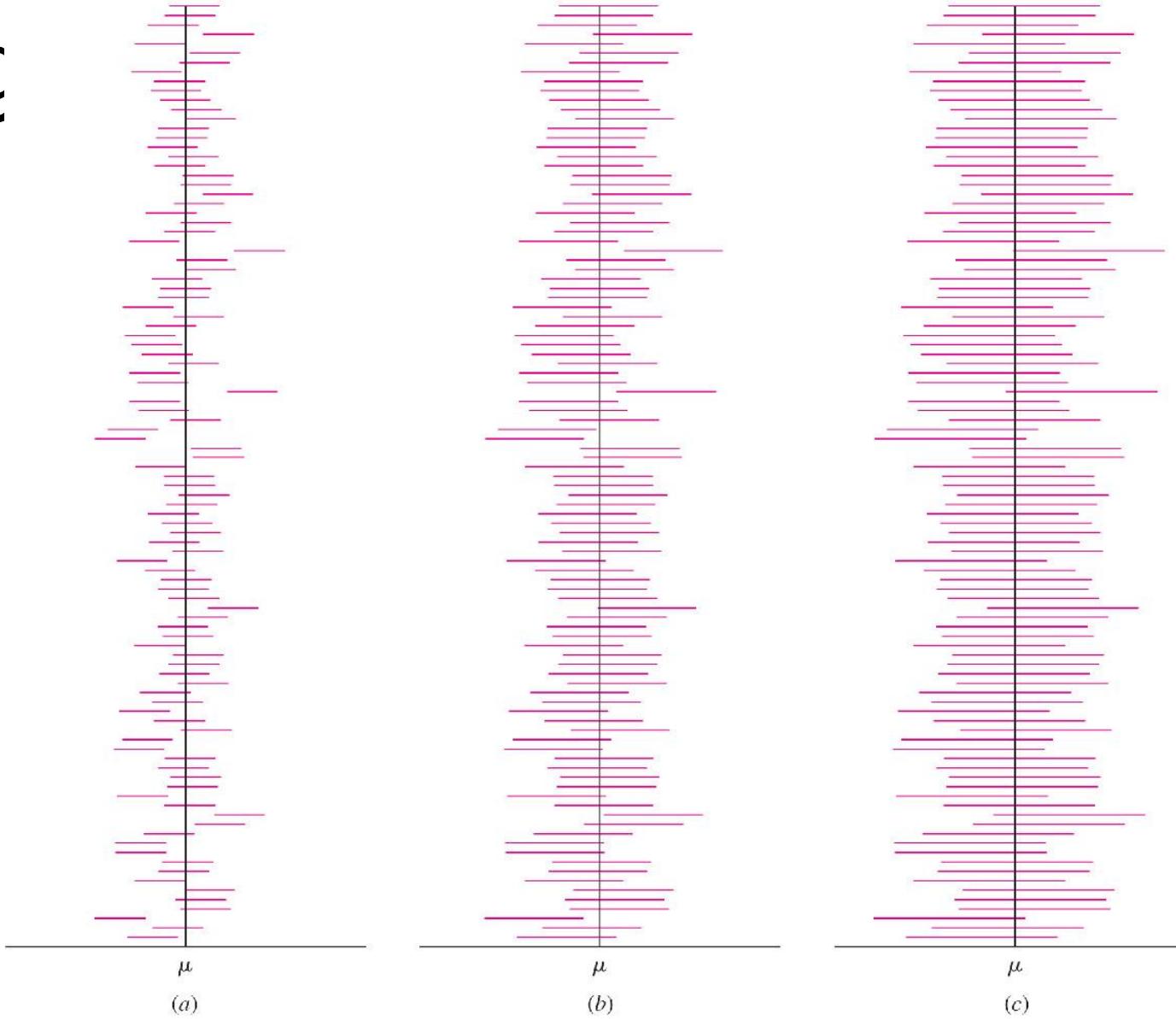
Outline

- ① Basic properties of confidence intervals
- ② Large-sample confidence intervals
- ③ Intervals based on a normal population distribution
- ④ Confidence intervals for the variance and standard deviation



Fig

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display



$$\bar{x} \pm 1.0 \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm 2.576 \frac{s}{\sqrt{n}}$$

Confidence interval when σ known

- This is an artificial case – for illustration only
- We assume that the population is normal and σ is known. Then

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

will have a standard normal distribution

- Hence

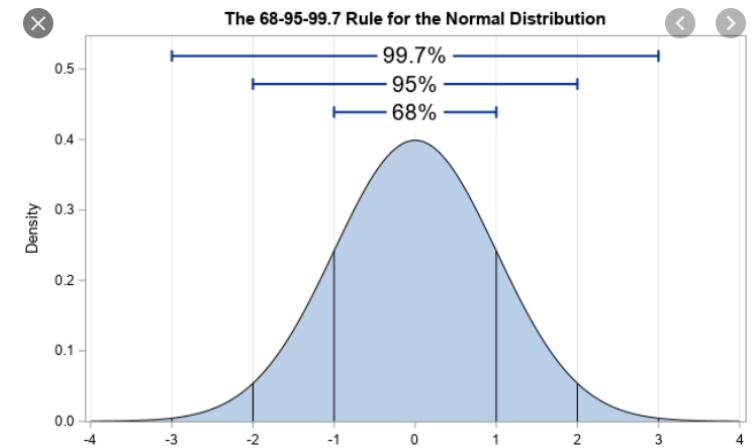
$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \quad (1)$$

- Now let's manipulate the inequalities inside the parentheses of (1)

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

$$-\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$



Properties of sample mean and sample sum

- Let X_1, X_2, \dots, X_n be a random sample from a distribution with mean μ and standard deviation σ . Then

$$\textcircled{1} \quad E[\bar{X}] = \mu_{\bar{X}} = \mu$$

$$E[\bar{X}] = E\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \frac{1}{n}[E(X_1) + E(X_2) + \dots + E(X_n)] = \frac{1}{n}n\mu = \mu$$

$$\textcircled{2} \quad V[\bar{X}] = \sigma_{\bar{X}}^2 = \sigma^2/n \text{ and } \sigma_{\bar{X}} = \sigma/\sqrt{n}$$

$$V[\bar{X}] = V\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \frac{1}{n^2}[V(X_1) + V(X_2) + \dots + V(X_n)] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

- If the original distribution of the X_i s is normal, then the distribution of \bar{X} and T_n is also normal.

Confidence interval when σ known

- This implies that

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- We construct the 95% confidence interval on μ as

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Interpretation

If we were repeatedly to sample from the distribution, about 95% of the intervals calculated in this way would cover the true mean μ .

The calculation of the interval from one sample would be like

```
> (samp <- rnorm(10, mean=6.3, sd=0.75))  
[1] 5.92 6.75 7.39 5.41 5.92 ...  
  
> mean(samp)  
[1] 6.393319  
  
> mean(samp)+c(lower=-1, upper=1) * 1.96 * 0.75/sqrt(10)  
lower          upper  
5.928464    6.858174
```

```
> (samp <- matrix(rnorm(50*10, mean=6.3, sd=0.75), nr=10))[1:5]
```

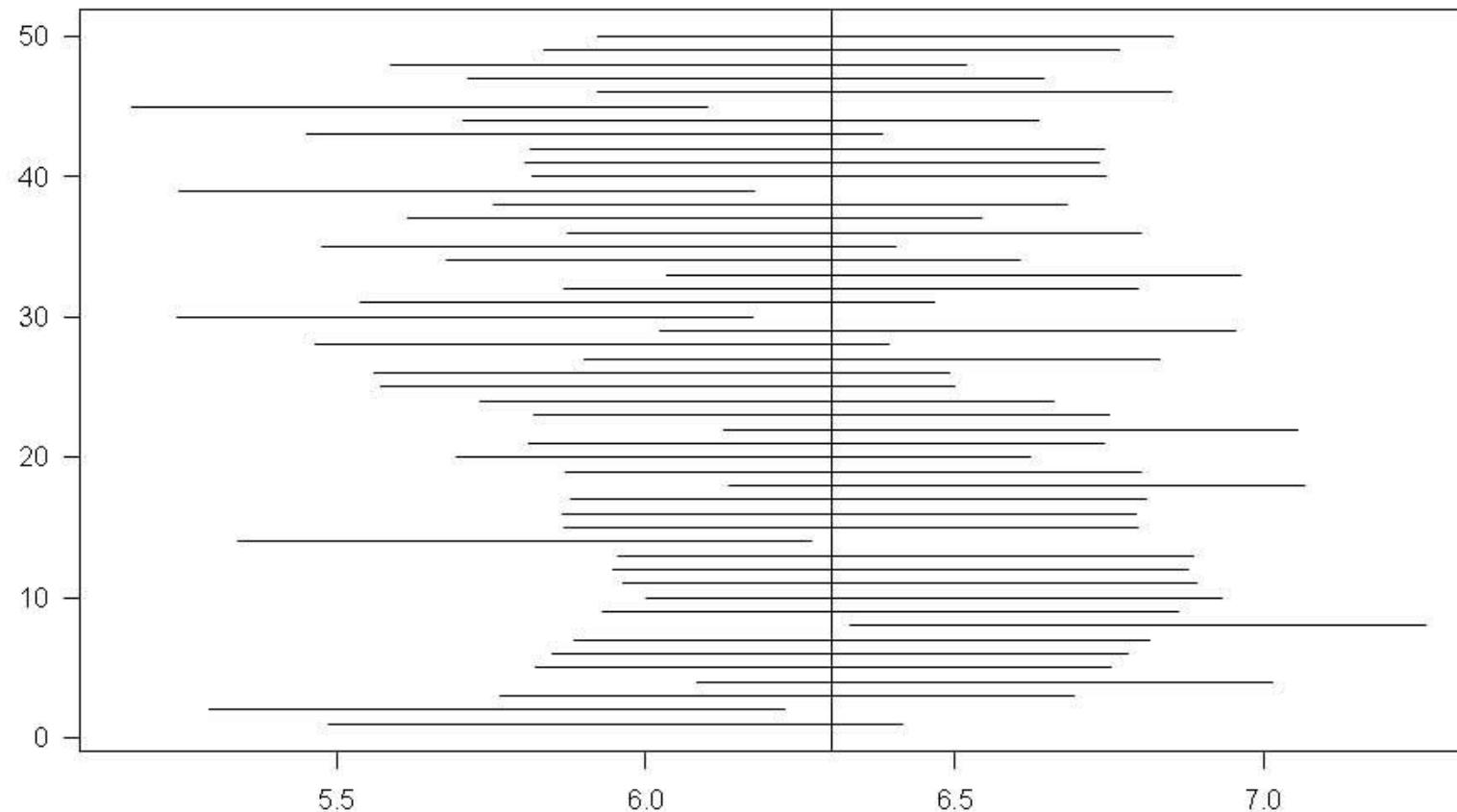
	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	6.870958	6.527197	6.899196	5.908721	5.917030
[2,]	4.479679	5.197282	8.139894	7.884092	5.491819
[3,]	5.816830	5.630525	6.060385	7.223066	6.430127
[4,]	4.581629	6.245015	5.778225	5.055725	6.564454
[5,]	7.215877	6.707928	6.590626	5.660643	5.993251
[6,]	6.281721	6.443767	6.730829	6.537824	5.372925
[7,]	5.358579	6.277609	6.782129	6.280670	5.683414
[8,]	7.109161	6.567141	7.609636	6.417673	5.638521
[9,]	5.598220	6.940539	5.635173	5.709908	6.340991
[10,]	7.847204	5.194830	6.520720	6.691987	5.799198

```
> (lims <- outer(c(lower=-1, upper=1) * 1.96 * 0.75/sqrt(10), colMeans(samp), "+"))[1:5]
```

	[,1]	[,2]	[,3]	[,4]	[,5]
lower	5.651131	5.708328	6.209826	5.872176	5.458318
upper	6.580841	6.638038	7.139536	6.801886	6.388028

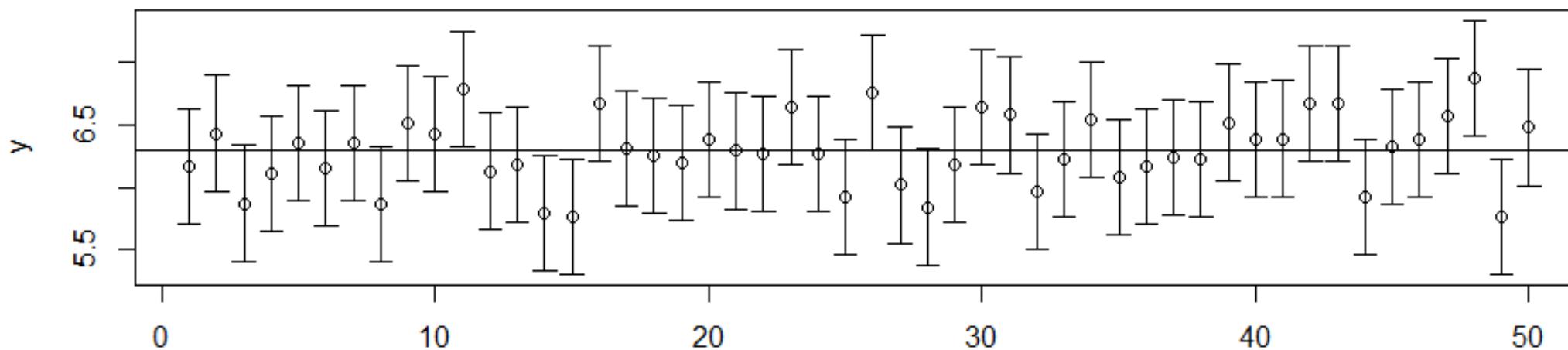
Plots of simulated confidence intervals

In this case only 44/50 or 88% of the intervals cover $\mu = 6.3$.



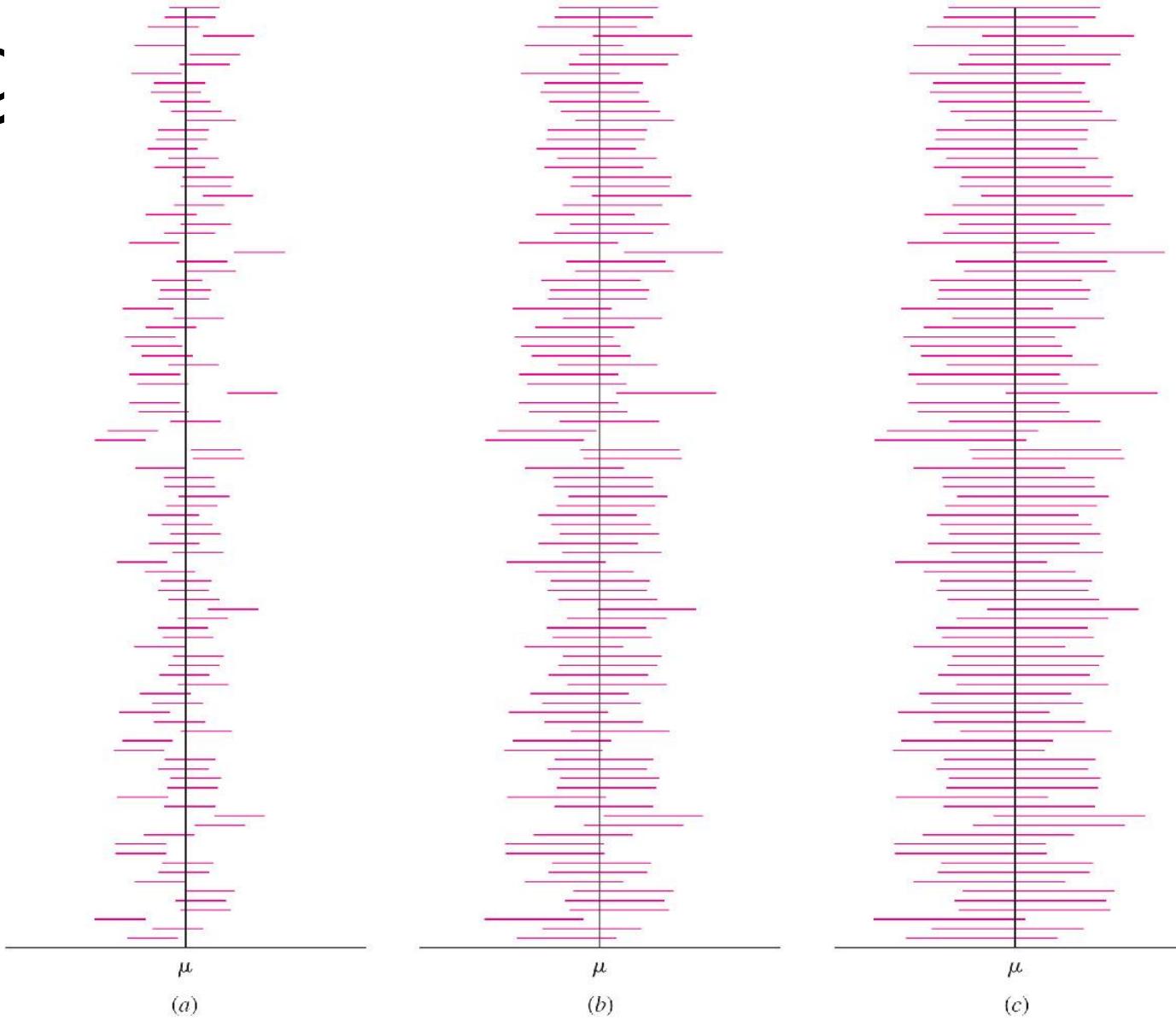
```
> library(plotrix)
> y <- colMeans(samp)
> err <- 1.96 * 0.75/sqrt(10)
> plotCI(1:50,y,err,main="Basic plotCI")
> abline(h=6.3)
```

Basic plotCI



Fig

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display



$$\bar{x} \pm 1.0 \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm 2.576 \frac{s}{\sqrt{n}}$$

Other levels of confidence

The multiplier 1.96 is determined from probabilities of the standard normal curve.
In general, a $100(1-\alpha)\%$ confidence interval on μ (when σ is known) is

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

Some common values of the multiplier are

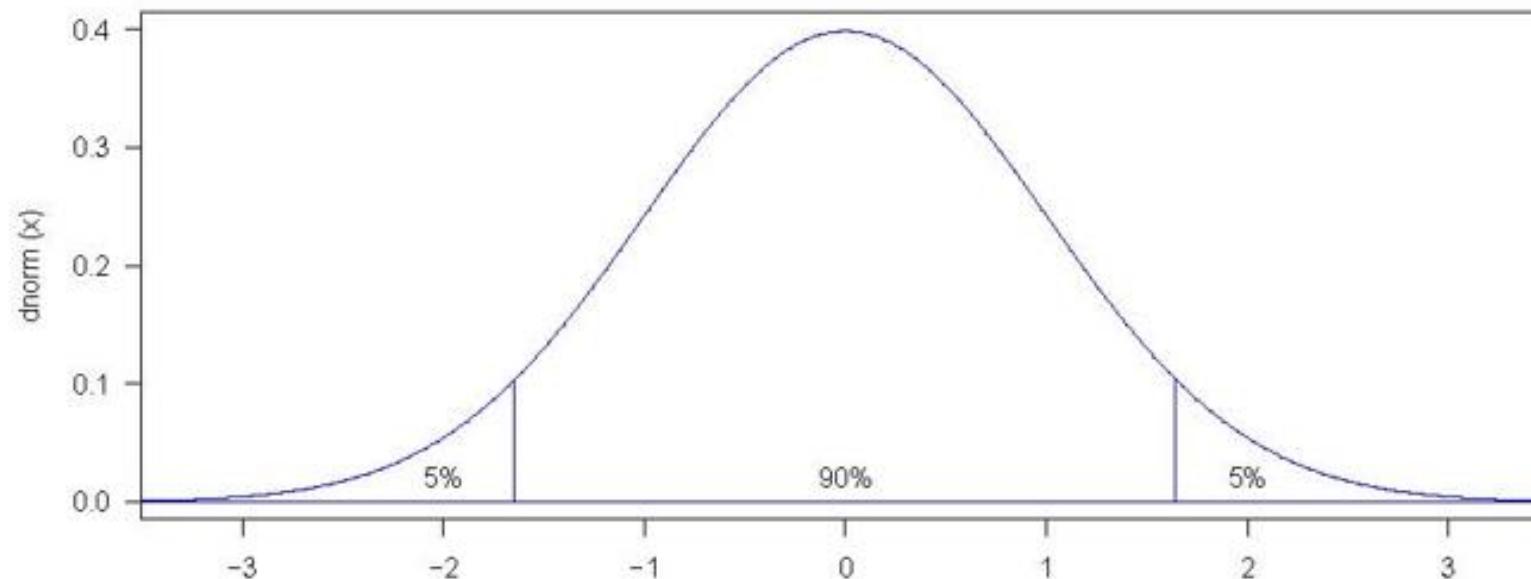
$100(1-\alpha)\%$	α	$\alpha/2$	$z_{\alpha/2}$
80%	0.20	0.10	1.282
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = 1 - \alpha \quad \rightarrow \quad P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Graph showing critical values

For a 90% confidence interval we use $z_{\alpha/2} = z_{0.05} = 1.645$.

When the standard normal curve is divided at -1.645 and 1.645, it has 5% of the area in the left tail, 90% in the middle, and 5% in the right tail.



Confidence Level, Precision, and Sample Size

Level : Increasing the confidence level requires increasing the multiplier $z_{\alpha/2}$. The only "100%" confidence interval is $(-\infty, \infty)$.

Sample size : Occasionally we know (or have an estimate of) σ and we want to determine the sample size required to get a given width for a confidence interval.

For example, if $\sigma = 25$, what n is required to have a 95% CI with width at most 10?

$$\bar{x} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm 1.96 \frac{25}{\sqrt{n}}$$

$$10 = 2 \cdot (1.96) \left(\frac{25}{\sqrt{n}} \right) \rightarrow n = \left[2 \cdot (1.96) \left(\frac{25}{10} \right) \right]^2 = 96$$

The general expression for the n to give a width of w in a $100(1-\alpha)\%$ CI is

$$w = 2 \cdot z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \quad n = \left(2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{w} \right)^2.$$

Large sample confidence interval when σ unknown

- 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

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

is a large-sample confidence interval for μ with confidence level approximately $100(1 - \alpha)\%$.

This formula is valid regardless of the shape of the population distribution.

- Central limit theorem : $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \rightarrow N(0, 1)$ as $n \rightarrow \infty$

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < z_{\alpha/2}\right) = 1 - \alpha \quad \rightarrow \quad P\left(\bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Large sample confidence interval

Example : xmp07.06 data

```
> str(xmp07.06)
'data.frame': 48 obs. of 1 variable:
 $ Voltage: int 62 50 53 57 41 53 55 61 59 64 ...
```

```
> with(xmp07.06, summary(Voltage))
   Min. 1st Qu. Median  Mean 3rd Qu. Max.
41.00  51.50  55.00  54.71  57.00  68.00
```

```
> with(xmp07.06, sd(Voltage))
```

```
[1] 5.230672
```

```
> qqmath(~Voltage, xmp07.06)
```

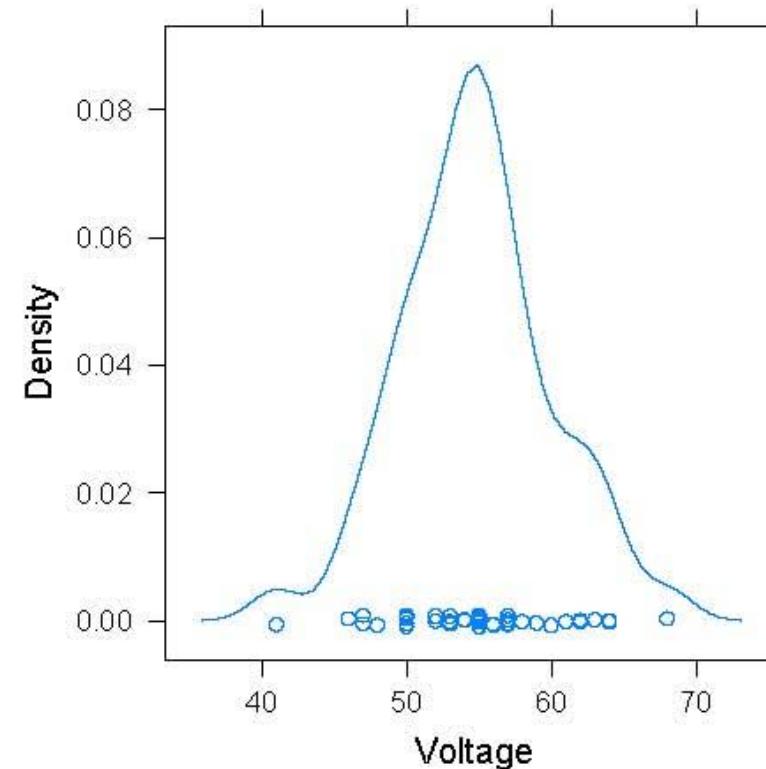
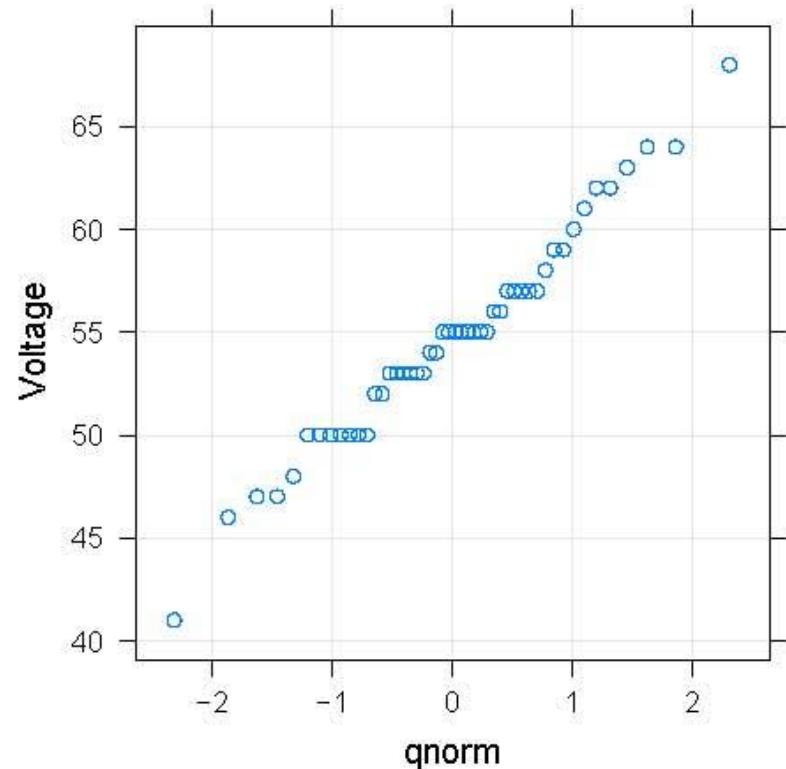
$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 54.708 \pm 1.96 \times \frac{5.2307}{\sqrt{48}} = (53.228, 56.188)$$

➤ In R, there is no "z interval". For large samples you can use the t.test function to get the interval

Large sample confidence interval

```
> with(xmp07.06, mean(Voltage) + c(lower = -1, upper = +1) *  
+     1.96 * sd(Voltage)/sqrt(length(Voltage)))
```

lower	upper
53.22857	56.18810



```
> with(xmp07.06, t.test(Voltage))
```

One Sample t-test

data: Voltage

t = 72.463, df = 47, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0 ($H_0 : \mu = 0$ vs $H_a : \mu \neq 0$)

95 percent confidence interval:

53.18950 56.22716

sample estimates:

mean of x

54.70833



$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} = \bar{x} \pm t_{0.025, 47} \frac{s}{\sqrt{n}} = 54.708 \pm 2.012 \times \frac{5.2307}{\sqrt{48}} = (53.189, 56.227)$$

> qt(0.975, 47) # 2.012 (when $X \sim N(\mu, \sigma^2)$, $T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$)

($z_{0.05}$ is computed by qnorm(0.95))

Confidence intervals on proportions

Distribution : If we have a random sample of binary data (i.e. yes/no data, success/failure data, etc.), the distribution of the number of successes would be binomial.

We write $\hat{P} = \frac{X}{n}$. The standard error of \hat{p} is $\sqrt{\hat{p}(1 - \hat{p})/n}$.

Large sample interval : The large-sample confidence interval on p uses the standard error

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n} \quad (7.11)$$

Adjusted interval : The adjustment given in the text is

$$\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n}}{1 + z_{\alpha/2}^2/n} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}. \quad (7.10)$$

A Confidence interval for a population proportions

- The natural estimator of p is $\hat{p} = \frac{X}{n}$, the sample fraction of successes.
- Since \hat{p} is just X multiplied by the constant $1/n$, \hat{p} also has approximately a normal distribution.

$$X = X_1 + X_2 + \cdots + X_n \text{ where } X_i \sim \text{Bernoulli}(p)$$

Central limit theorem : sample mean of Bernoulli r.v. $\hat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right), n \rightarrow \infty$

- As we know, $E(\hat{p}) = p$ (unbiasedness) and $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.
- The standard deviation $\sigma_{\hat{p}}$ involves the unknown parameter p .
- Standardizing \hat{p} by subtracting p and dividing by $\sigma_{\hat{p}}$ then implies that

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

A Confidence interval for a population proportions

$$P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha \quad \rightarrow \text{Solve for } \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} = z_{\alpha/2}$$

$$\hat{p} - p = z_{\alpha/2} \sqrt{p(1-p)/n} \quad \rightarrow \quad \hat{p}^2 - 2p\hat{p} + p^2 = z_{\alpha/2}^2 \{p(1-p)/n\}$$

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

$$(n + z_{\alpha/2}^2)p^2 - \left(z_{\frac{\alpha}{2}}^2 + 2n\hat{p}\right)p + n\hat{p}^2 = 0 \quad (\text{roots of quadratic equation } a x^2 + bx + c = 0 : x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a})$$

$$p = \frac{z_{\frac{\alpha}{2}}^2 + 2n\hat{p} \pm \sqrt{\left(z_{\frac{\alpha}{2}}^2 + 2n\hat{p}\right)^2 - 4(n + z_{\alpha/2}^2)n\hat{p}^2}}{2(n + z_{\alpha/2}^2)} \quad p = \frac{z_{\frac{\alpha}{2}}^2 + 2n\hat{p} \pm \sqrt{\left(z_{\frac{\alpha}{2}}^2\right)^2 + 4n\hat{p}z_{\frac{\alpha}{2}}^2 + 4n^2\hat{p}^2 - 4n^2\hat{p}^2 - 4n\hat{p}^2z_{\frac{\alpha}{2}}^2}}{2(n + z_{\alpha/2}^2)}$$

$$p = \frac{z_{\frac{\alpha}{2}}^2 + 2n\hat{p} \pm \sqrt{\left(z_{\frac{\alpha}{2}}^2\right)^2 + 4n\hat{p}z_{\frac{\alpha}{2}}^2 - 4n\hat{p}^2z_{\frac{\alpha}{2}}^2}}{2(n + z_{\alpha/2}^2)} = \frac{z_{\frac{\alpha}{2}}^2 + 2n\hat{p} \pm z_{\alpha/2} \sqrt{\left(z_{\frac{\alpha}{2}}^2\right)^2 + 4n\hat{p} - 4n\hat{p}^2}}{2(n + z_{\alpha/2}^2)} = \frac{\hat{p} + z_{\frac{\alpha}{2}}^2/2n}{1 + z_{\alpha/2}^2/n} \pm z_{\alpha/2} \frac{\sqrt{\left(z_{\frac{\alpha}{2}}^2\right)^2 + 4n\hat{p} - 4n\hat{p}^2}}{\sqrt{\frac{\left(z_{\frac{\alpha}{2}}^2\right)^2 + 4n\hat{p} - 4n\hat{p}^2}{4n^2}}}$$

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

Confidence intervals on proportions : Example 7.8

An article in J. Testing and The Eval. had 16 successes in 48 trials.

The 95% confidence intervals from the text are (0.217, 0.474) (with correction by eqn (7.10)) and

(0.200, 0.466) (without correction by eqn (7.11))

```
> prop.test(16, 48)
```

1-sample proportions test with continuity correction

data: 16 out of 48, null probability 0.5

X-squared = 4.6875, df = 1, p-value = 0.03038

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

0.2080794 0.4851357

0.2080794 < (0.217 < 0.474) < 0.4851357

sample estimates:

0.2080794 < (0.200 < 0.466) < 0.4851357

Confidence intervals on proportions : Example 7.8

```
> binom.test(16, 48)
```

Exact binomial test

data: 16 and 48

number of successes = 16, number of trials = 48, p-value = 0.0293

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.2039597 0.4841083

0.2039597 < (0.217 < 0.474) < 0.4841083

sample estimates:

0.2039597 < (0.200 < 0.466) < 0.4841083

probability of success

0.3333333

One-sided intervals

Occasionally we are only interested in bounding a parameter estimate above or bounding below. The main difference is that we use z_α instead of $z_{\alpha/2}$ as the multiplier.

ex) mean life time of a tire: $\mu > 70000$ (km)

The one-sided intervals are

$$\mu < \bar{x} + z_\alpha \frac{s}{\sqrt{n}} \quad \text{or} \quad \mu > \bar{x} - z_\alpha \frac{s}{\sqrt{n}} \quad (\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}})$$

$$P\left(\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} < z_\alpha\right) = 1 - \alpha$$

$$\bar{X} - \mu < z_\alpha \cdot \frac{s}{\sqrt{n}}$$

$$-\mu < -\bar{X} + z_\alpha \cdot \frac{s}{\sqrt{n}}$$

$$\bar{X} - z_\alpha \cdot \frac{s}{\sqrt{n}} < \mu \quad \rightarrow \quad P\left(\bar{X} - z_\alpha \cdot \frac{s}{\sqrt{n}} < \mu\right) = 1 - \alpha$$

One-sided intervals

```
> with(xmp07.06, t.test(Voltage, alt = "less"))
```

One Sample t-test

data: Voltage

t = 72.463, df = 47, p-value = 1

alternative hypothesis: true mean is less than 0

95 percent confidence interval:

-Inf 55.97514

sample estimates:

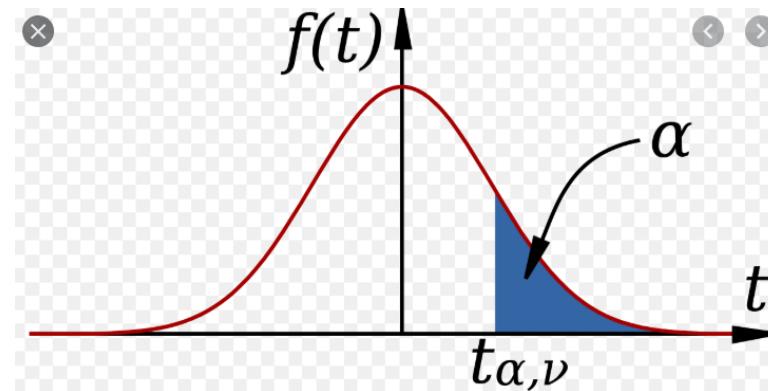
mean of x

54.70833

$$\bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}} = \bar{x} + t_{0.05, 47} \frac{s}{\sqrt{n}} = 54.708 + 1.678 \times \frac{5.2307}{\sqrt{48}} = 55.97514$$

$$> qt(0.95, 47) \# 1.678 \quad (\text{when } X \sim N(\mu, \sigma^2), T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1})$$

($z_{0.05}$ is computed by qnorm(0.95))



One-sided intervals

```
> with(xmp07.06, t.test(Voltage, alt = "greater"))
```

One Sample t-test

data: Voltage

t = 72.463, df = 47, p-value < 2.2e-16

alternative hypothesis: true mean is greater than 0

95 percent confidence interval:

53.44153 Inf

sample estimates:

mean of x

54.70833

$$\bar{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}} = \bar{x} - t_{0.05,47} \frac{s}{\sqrt{n}} = 54.708 - 1.678 \times \frac{5.2307}{\sqrt{48}} = 54.44153$$

```
> qt(0.95, 47) # 1.678
```

(when $X \sim N(\mu, \sigma^2)$, $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$)

($z_{0.05}$ is computed by `qnorm(0.95)`)

Intervals based on a Normal Distribution

Assumption : When working with a small sample we must make additional assumptions on the distribution to make up for our lack of information about σ .

We assume the X_i 's are from a normal distribution.

Check this with a normal probability plot.

T distribution : Assuming that we have a random sample from a normal distribution

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} : \nu \text{ is the degrees of freedom}$$

$$\left(\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, \Gamma(n) = (n-1)!, \Gamma(1) = 1, \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \right)$$

Intervals based on a Normal Population Distribution

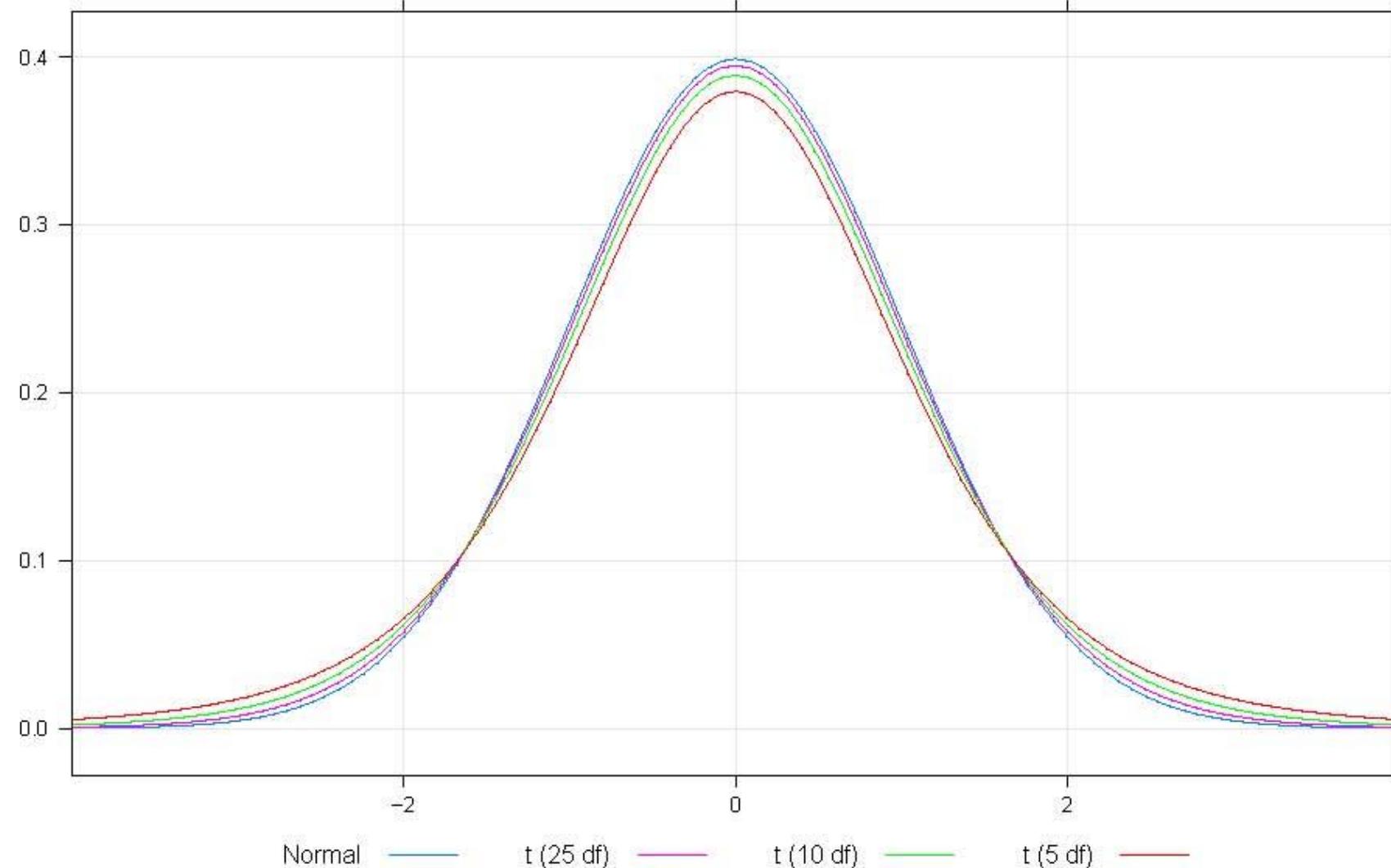
- The key result underlying the interval in earlier section was that for large n ,
the r.v. $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ has approximately a standard normal distribution.
- When n is small, S is no longer likely to be close to σ , so the variability in the distribution of Z arises from randomness in both the numerator and the denominator.
- This implies that the probability distribution of $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ will be more spread out than the standard normal distribution.

Properties of the T distribution

- ① Each t_ν curve is bell-shaped and centered at 0
- ② Each t_ν curves is more spread out than the standard normal.
- ③ As $\nu \uparrow$, the spread of t_ν decreases
- ④ As $\nu \uparrow \infty$, t_ν approaches the standard normal.

Graphical comparison of t_ν and z

Comparison of t densities and a standard normal density



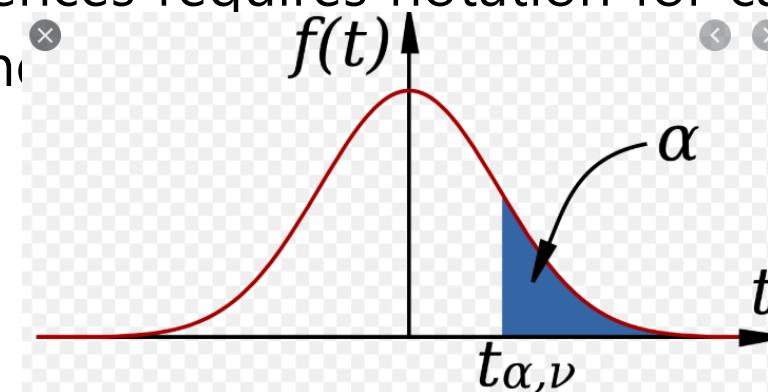
Properties of the t distribution

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

- The number of *df* for T is $n - 1$.

Although S is based on the n deviations $X_1 - \bar{X}, \dots, X_n - \bar{X}$, the fact that $\sum(X_i - \bar{X}) = 0$ implies that only $n - 1$ of these are “freely determined”

- The number of *df* for a *t* variable is the number of freely determined deviations on which the estimated standard deviation in the denominator of T is based.
- The use of *t* distribution in making inferences requires notation for capturing *t*-curve tail areas t_α analogous to z_α , for thi



T critical values

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

Definition : A **t critical value**, denoted $t_{\alpha,\nu}$, is the point on the t_ν curve with probability α to the right.

Text Critical values for selected ν and α are given in table A.8, A-5 (on inside back cover)

R Use the qt function with lower=FALSE

```
> qt(0.05, df=15, lower=FALSE)
```

```
[1] 1.75305
```

or

```
> qt(0.95, df=15) #1.75305
```

The first argument is α (or $1 - \alpha$). The df argument is ν .

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

Definition : A **t critical value**, denoted $t_{\alpha,\nu}$, is the point on the t_ν curve with probability α to the right.

Text Critical values for selected ν and α are given in table A.8, A-5 (on inside back cover)

R Use the qt function with lower=FALSE

```
> qt(0.05, df=15, lower=FALSE)
```

```
[1] 1.75305
```

or

```
> qt(0.95, df=15) #1.75305
```

The first argument is α (or $1 - \alpha$). The df argument is ν .

t Table

cum. prob	<i>t</i> . _{.50}	<i>t</i> . _{.75}	<i>t</i> . _{.80}	<i>t</i> . _{.85}	<i>t</i> . _{.90}	<i>t</i> . _{.95}	<i>t</i> . _{.975}	<i>t</i> . _{.99}	<i>t</i> . _{.995}	<i>t</i> . _{.999}	<i>t</i> . _{.9995}
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850

General confidence interval on μ

Two-sided The general two-sided interval is

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \quad \text{cf) } \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

This is provided by `t.test` in *R*.

Use the argument `conf.level` to get a level other than the default of 95%.

One-sided An upper confidence bound is provided by

$$\bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}} \quad \text{cf) } \bar{x} + z_{\alpha} \frac{s}{\sqrt{n}}$$

Use `alt="less"` in `t.test` to get this interval.

A lower bound is provided by `alt="greater"` and calculated as

$$\bar{x} - t_{\alpha, n-1} s / \sqrt{n} \quad \text{cf) } \bar{x} - z_{\alpha} \frac{s}{\sqrt{n}}$$

Confidence interval when σ is unknown

Two-sided interval :

$$\begin{aligned} P\left(-t_{\frac{\alpha}{2}, n-1} < T = \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2, n-1}\right) &= 1 - \alpha \\ -t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} < \bar{X} - \mu &< t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \\ -\bar{X} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} < -\mu &< -\bar{X} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \\ \bar{X} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} < \mu &< \bar{X} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \end{aligned}$$

One-sided interval :

$$\begin{aligned} P\left(T = \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha, n-1}\right) &= 1 - \alpha \\ \bar{X} - \mu &< t_{\alpha, n-1} \cdot \frac{s}{\sqrt{n}} \\ -\mu &< -\bar{X} + t_{\alpha, n-1} \cdot \frac{s}{\sqrt{n}} \\ \bar{X} - t_{\alpha, n-1} \cdot \frac{s}{\sqrt{n}} &< \mu \end{aligned}$$

```
> with(xmp07.06, t.test(Voltage))
```

One Sample t-test

data: Voltage

t = 72.463, df = 47, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0 ($H_0 : \mu = 0$ vs $H_a : \mu \neq 0$)

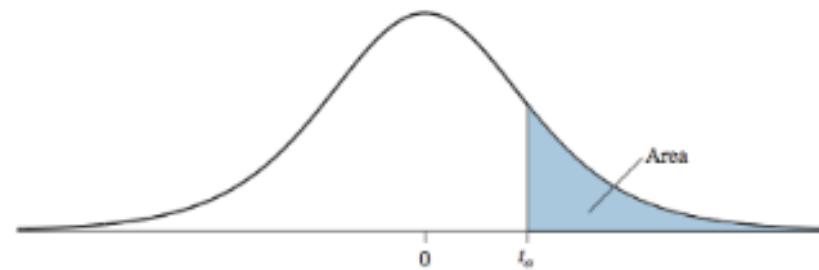
95 percent confidence interval:

53.18950 56.22716

sample estimates:

mean of x

54.70833



$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} = \bar{x} \pm t_{0.025, 47} \frac{s}{\sqrt{n}} = 54.708 \pm 2.012 \times \frac{5.2307}{\sqrt{48}} = (53.189, 56.227)$$

> qt(0.975, 47) # 2.012 (when $X \sim N(\mu, \sigma^2)$, $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$)

($z_{0.05}$ is computed by qnorm(0.95))

The One-Sample t Confidence Interval : Example 7.11

Data on the modulus of rupture of composite beams designed to add value to low-grade sweetgum lumber is given in an article in J. of Bridge Engr.

6807.99, 7637.06, 6663.28, 6165.03, 6991.41, 6992.23, 6981.46, 7569.75,
7437.88, 6872.39, 7663.18, 6032.28, 6906.04, 6617.17, 6984.12, 7093.71,
7659.50, 7378.61, 7295.54, 6702.76, 7440.17, 8053.26, 8284.75, 7347.95,
7422.69, 7886.87, 6316.67, 7713.65, 7503.33, 7674.99

$$95\% \text{ CI} : \bar{x} \pm t_{0.025,29} \frac{s}{\sqrt{30}}$$



Sweetgum (*Liquidambar styraciflua*)

- `modulus <- c(6807.99, 7637.06, 6663.28, 6165.03, 6991.41, 6992.23, 6981.46, 7569.75, 7437.88, 6872.39, 7663.18, 6032.28, 6906.04, 6617.17, 6984.12, 7093.71, 7659.50, 7378.61, 7295.54, 6702.76, 7440.17, 8053.26, 8284.75, 7347.95, 7422.69, 7886.87, 6316.67, 7713.65, 7503.33, 7674.99)`

95% confidence interval :

```
> mean(modulus) + c(lower=-1, upper=+1) *  
qt(0.975,29)*sd(modulus)/sqrt(length(modulus))
```

```
>
```

lower	upper
7000.230	7406.152

```
> t.test(modulus)
```

95 percent confidence interval:

7000.230 7406.152

Example of confidence interval on μ : Example 7.14

Data on the modulus of elasticity of Scotch pine lumber specimens is given in an article in J. of Testing and Eval.

```
> str(xmp07.11)
```

'data.frame': 16 obs. of 1 variable:

\$ Elasticity: int 10490 16620 17300 15480 12970 17260 13400 13900 13630 13260 ...

```
> elasticity <- xmp07.11$Elasticity
```

```
> mean(elasticity) + c(lower=-1, upper=+1) * t(0.975,15)*sd(elasticity)/sqrt(length(elasticity))
```

lower	upper
-------	-------

13437.11	15627.89
----------	----------

```
> t.test(xmp07.11$Elasticity)
```

$$95\% \text{ CI} : \bar{x} \pm t_{0.025,15} \frac{s}{\sqrt{16}}$$

Prediction

A prediction interval is a confidence interval on the value of a future response(observation).

This interval incorporates the variability in the estimate of the unknown mean μ and the variability in the response itself.

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

This type of interval is much less common than the confidence interval on μ .

We have a random sample X_1, X_2, \dots, X_n from a normal distribution, and wish to predict the value of X_{n+1} , a single future observation.

$$E(\bar{X} - X_{n+1}) = E(\bar{X}) - E(X_{n+1}) = \mu - \mu = 0$$

$$V(\bar{X} - X_{n+1}) = V(\bar{X}) + V(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2(1 + \frac{1}{n})$$

$$T = \frac{\bar{X} - X_{n+1} - 0}{s\sqrt{1+1/n}} \sim t \text{ distribution with } n-1 \text{ df}$$

$$P\left(-t_{\frac{\alpha}{2}, n-1} < T = \frac{\bar{X} - X_{n+1}}{s\sqrt{1+1/n}} < t_{\alpha/2, n-1}\right) = 1 - \alpha$$

Example 7.12, 7.13

Consider the following sample of fat content (in percentage) of $n = 10$ randomly selected hot dogs .

25.2 21.3 22.8 17.0 29.8 21.0 25.5 16.0 20.9 19.5

Assume that these were selected from a normal population distribution,
95% CI :

$$\bar{x} \pm t_{0.025,9} \frac{s}{\sqrt{n}} = 21.90 \pm 2.262 \frac{4.134}{\sqrt{10}} = 21.90 \pm 2.96 = (18.94, 24.86)$$

95% PI :

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s \sqrt{1 + \frac{1}{n}} = 21.90 \pm (2.262)(4.134) \sqrt{1 + \frac{1}{10}} = 21.90 \pm 9.81 = (12.09, 31.71)$$

Confidence intervals for Variance

- A confidence interval on the variance σ^2 or the standard deviation σ of a normal population can be derived from the distribution of S^2

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \left(S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}, (n-1)S^2 = \sum(X_i - \bar{X})^2 \right)$$

- Because the χ^2 distribution is not symmetric, we need the critical values $\chi_{1-\alpha/2, n-1}^2$ and $\chi_{\alpha/2, n-1}^2$
- Take square roots of the interval on the variance to get an interval on the standard deviation.
- These intervals are uncommon.

If Z_1, \dots, Z_k are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^k Z_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom.

This is usually denoted as

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi_k^2$$

$$\frac{\sum(X_i - \mu)^2}{\sigma^2} = \sum Z_i^2 \sim \chi_k^2, \quad \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

➤ Gamma distribution

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

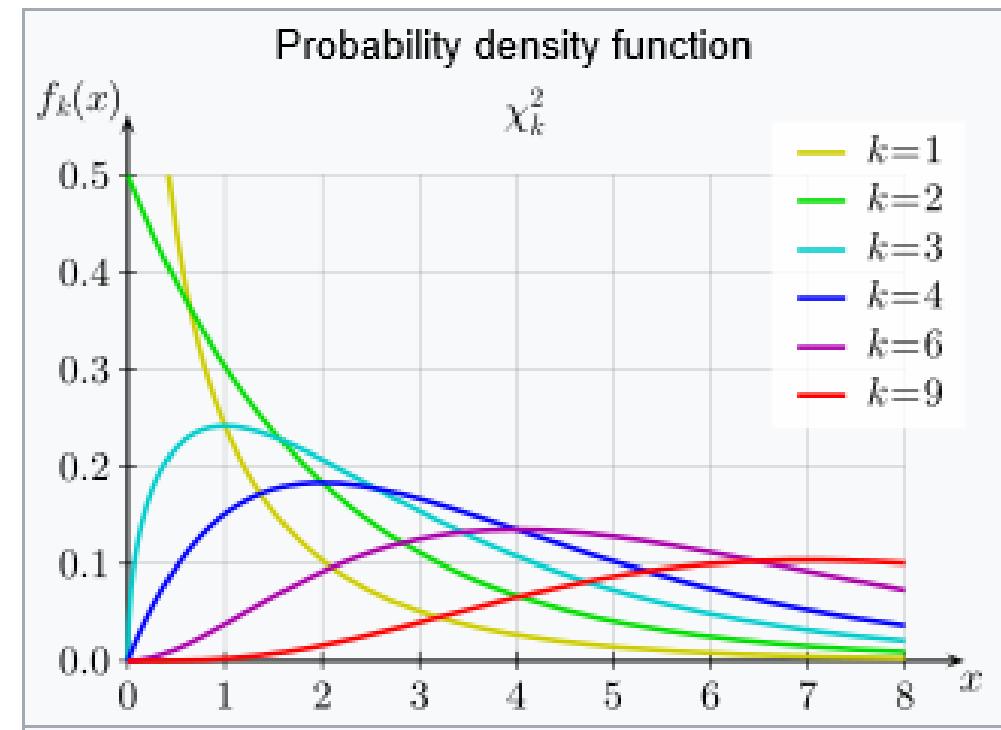
➤ Chi-squared distribution :

shape parameter : $\alpha = \frac{k}{2}$

scale parameter : $\beta = 2$

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

chi-squared



Confidence intervals of σ^2

From the theorem,

$$P \left(\chi_{1-\alpha/2, n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2, n-1}^2 \right) = 1 - \alpha$$

We get the inequalities

$$\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}$$

A $100(1 - \alpha)\%$ confidence interval for the variance σ^2 of a normal population is

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2}, \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2} \right)$$

Example 7.15

The accompanying data is the breakdown voltage of electrically stressed circuits. What is 95% CI for σ^2 ?

1470 1510 1690 1740 1900 2000 2030 2100 2190
2200 2290 2380 2390 2480 2500 2580 2700

(Solution) $s^2 = 137324.3$, $df=n-1=16$, $\chi^2_{0.975,16} = 6.908$, $\chi^2_{0.025,16} = 28.845$

The CI is

$$\left(\frac{16(137324.3)}{28.845}, \frac{16(137324.3)}{6.908} \right) = (76,172.3, 318,064.4)$$

```
> voltage <- c(1470, 1510, 1690, 1740, 1900, 2000, 2030, 2100, 2190,  
+ 2200, 2290, 2380, 2390, 2480, 2500, 2580, 2700)  
  
> var(voltage)  
[1] 137324.3  
  
> 16*var(voltage)/qchisq(0.975, 16)  
[1] 76171.31  
  
> 16*var(voltage)/qchisq(0.025, 16)  
[1] 318079.8
```

Chapter 8 – Single sample hypothesis tests

Outline

- ① Hypotheses and test procedures
- ② Tests About a Population Mean
- ③ Tests Concerning a Population Proportion
- ④ P-values

Statistical Hypotheses

- A statistical hypothesis is a claim about the value of one or more population parameters.
- **Null Hypothesis** This is the claim that is initially assumed to be true (the prior belief claim) : H_0 .
- **Alternative Hypothesis** : This is the assertion that is contradictory to H_0 . : H_a .
- The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false.

If the sample does not strongly contradict H_0 , we will continue to believe in the plausibility of the null hypothesis.

The two possible conclusions from a hypothesis-testing analysis are,

reject H_0 or fail to reject H_0 .

Smoking example

- Null hypothesis (H_0) : Smoking does not cause the lung cancer
Alternative hypothesis (H_a) : Smoking causes the lung cancer.
- [\\$26.75M Award in Retrial Over Smoker's Death More Than Doubles Original Trial Verdict](#)

Posted by [Arlin Crisco](#) on Feb 18, 2020 4:27:29 PM

St. Petersburg, FL— A Florida state court jury awarded \$26.75 million to the family of a Florida smoker after finding the nation's two largest tobacco companies responsible for his cancer death. [Duignan v. R.J. Reynolds and Philip Morris, 13-010978-CI.](#)

Duignan, 42 when he died, smoked up to two packs of cigarettes a day for more than 25 years. His family contends Reynolds and Philip Morris's role in a conspiracy to hide the dangers of cigarettes hooked Duignan to nicotine and caused his fatal cancer.

\$26.75M Award in Retrial Over Smoker's Death More Than Doubles Original Trial Verdict

Posted by *Arlin Crisco* on Feb 18, 2020 4:27:29 PM

 Tweet

 Share

 Like

Share



Stock image.

St. Petersburg, FL— A Florida state court jury awarded \$26.75 million to the family of a Florida smoker after finding the nation's two largest tobacco companies responsible for his cancer death. *Duignan v. R.J. Reynolds and Philip Morris, 13-010978-CI.*

Court example

- Null hypothesis (H_0) : Defendant is assumed to be innocent.
Alternative Hypothesis (H_a) : Defendant is guilty.
- When there exists enough evidence that the defendant is guilty, he receives a verdict of guilty.
- Error probability :
 - Defendant receives a verdict of guilty when he is innocent.
This probability is kept very low.
 - Defendant receives a verdict of not guilty when he is guilty.
O. J. Simpson trial : former college and professional football star (defence team : dream team)
 - In criminal trial, found not guilty of the murder of his wife
 - In civil trial, found responsible for the death of his wife and fined \$33.5 million

Oct. 5-7, 1995 Gallup Poll
Simpson: Criminal Trial Verdict

	Verdict Right	Verdict Wrong	No Opinion
Total	47%	44%	9%
Whites	42%	49%	9%
Blacks	78%	10%	12%

1999 Gallup Poll
Did Simpson commit murder?

Definitely	36%
Probably	38%
Probably Not	15%
Definitely Not	6%
No Opinion	6%

Test Procedures

- A test procedure is specified by the following :
 1. A test statistic : a function of the sample data on which the decision (reject H_0 or do not reject H_0) is to be based.
 2. A rejection region : the set of all test statistic values for which H_0 will be rejected
- The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

Errors in Hypothesis Testing : Example 8.1

A type I error(α) consists of rejecting the null hypothesis H_0 when it is true.

A type II error(β) involves not rejecting H_0 when H_0 is false.

A certain type of automobile is known to have no visible damage 25% of the time in 10-mph crash tests.

A modified bumper design has been proposed in an effort to increase this percentage.

$$H_0 : p = 0.25 \text{ vs } H_a : p > 0.25.$$

The test will be based on an experiment involving $n=20$ independent crashes with cars of new design.

Consider the following test procedure:

Test statistic: X = the number of crashes with no visible damage

Rejection region: $R_8 = \{ 8, 9, 10, \dots, 19, 20 \}$; that is, reject H_0 if $x \geq 8$

*Type I + Type II Errors

Type I Error: Rejecting the null hypothesis when it is true.

Type 2 Error: Not rejecting the null hypothesis when it is false.

$$P(\text{type I error} / H_0 \text{ is true}) = \alpha$$

$$P(\text{type II error} / H_0 \text{ is false}) = \beta$$

$$P(\text{rejecting a false } H_0) = 1 - \beta$$

		H_0
		True
Reject H_0	True	\checkmark
	False	Type I Error
Fail to Reject H_0	\checkmark	Type II Error

Court example

- Null hypothesis (H_0) : Defendant is assumed to be innocent.
Alternative Hypothesis (H_a) : Defendant is guilty.
- When there exists enough evidence that the defendant is guilty, he receives a verdict of guilty.
- Error probability :
 - Defendant receives a verdict of guilty when he is innocent. : type I error
This probability is kept very low.
 - Defendant receives a verdict of not guilty when he is guilty. : type II error

Errors in Hypothesis Testing

When H_0 is true, X has a binomial probability distribution with $n = 20$ and $p = 0.25$. Then

$$\alpha = P(\text{type I error})$$

$$= P(H_0 \text{ is rejected when it is true})$$

$$= P(X \geq 8 \text{ when } X \sim \text{Bin}(n = 20, p = 0.25))$$

$$= 1 - B(7; 20, .25) = 0.102$$

$$1 - \sum_{x=0}^7 \binom{20}{x} 0.25^x (1 - 0.25)^{20-x}$$

$$> 1 - \text{pbinom}(7, 20, 0.25) \# 0.1018119$$

$$\beta(0.3) = P(\text{type II error when } p = 0.3)$$

$$= P(H_0 \text{ is not rejected when it is false because } p = 0.3)$$

$$= P(X \leq 7 \text{ when } X \sim \text{Bin}(20, .3))$$

$$= B(7; 20, .3) = 0.772$$

$$\sum_{x=0}^7 \binom{20}{x} 0.3^x (1 - 0.3)^{20-x}$$

$$> \text{pbinom}(7, 20, 0.3) \# 0.7722718$$

Errors in Hypothesis Testing

- Proposition

Suppose an experiment and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of α results in a larger value of β for any particular parameter value consistent with H_a .

- This proposition says that once the test statistic and n are fixed, there is no rejection region that will simultaneously make both α and all β 's small.
- We specify the largest value of α that can be tolerated. α is referred to as the **significance level** of the test.
- The corresponding test procedure is called a level α test (e.g., a level 0.05 test or a level 0.01 test).

Bumper example: Consider the following test procedure:

Test statistic: X = the number of crashes with no visible damage

Rejection region: $R_9 = \{9, 10, \dots, 19, 20\}$; that is, reject H_0 if $x \geq 9$

When H_0 is true, X has a binomial probability distribution with $n = 20$ and $p = 0.25$. Then

$$\begin{aligned}\alpha &= P(\text{type I error}) \\ &= P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq 9 \text{ when } X \sim \text{Bin}(n = 20, p = 0.25)) \\ &= 1 - B(8; 20, .25) = 0.04092517 \quad (\text{reduced from } 0.1018119)\end{aligned}$$

$$\begin{aligned}\beta(0.3) &= P(\text{type II error when } p = 0.3) \\ &= P(H_0 \text{ is not rejected when it is false because } p = 0.3) \\ &= P(X \leq 8 \text{ when } X \sim \text{Bin}(20, .3)) \\ &= B(8; 20, .3) = 0.8866685 \quad (\text{increased from } 0.7722718)\end{aligned}$$

Test about a population mean with Known σ

Let X_1, \dots, X_n represent a random sample from the normal distribution.

Null hypothesis : $H_0 : \mu = \mu_0$

Test statistic value : $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis Rejection Region for level α test

$$H_a : \mu > \mu_0 \quad z \geq z_\alpha \text{ (upper tailed)}$$

$$H_a : \mu < \mu_0 \quad z \leq -z_\alpha \text{ (lower tailed)}$$

$$H_a : \mu \neq \mu_0 \quad \text{either } z \geq z_{\frac{\alpha}{2}} \text{ or } z \leq -z_{\alpha/2} \text{ (two-tailed)}$$

Test about a population mean with Known σ

- Use of the following sequence of steps is recommended
- ① Identify the parameter of interest and describe it in the context of the problem situation.
 - ② Determine the null value and state the null hypothesis.
 - ③ State the appropriate alternative hypothesis.
 - ④ Give the formula of the test statistic.
 - ⑤ State the rejection region for the selected significance level .
 - ⑥ Compute any necessary sample quantities, substitute into the formula
 - ⑦ Decide whether H_0 should be rejected

Test about a population mean with known σ

A manufacturer of sprinkler systems claims that the true average system-activation temperature is 130°F.

A sample of $n = 9$ systems yields a sample average activation temperature of 131.08 °F.

If the distribution of activation times is normal with standard deviation 1.5 °F, does the data contradict the manufacturers claim at significance level $\alpha = 0.01$?

- ① Parameter of interest: μ = true average activation temperature.
- ② Null hypothesis: $H_0 : \mu = 130$ (null value : $\mu_0 = 130$).
- ③ Alternative hypothesis: $H_a : \mu \neq 130$ (a departure from the claimed value in either direction is of concern).

Test about a population mean with Known σ

- ④ Test statistic value: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{9}}$
- ⑤ We reject H_0 if either $z \geq 2.58 = z_{0.005}$ or $z \leq -2.58 = -z_{0.005}$:
- ⑥ Substituting $n = 9$ and $\bar{x} = 131.08$, we get

$$z = \frac{\bar{x} - 130}{1.5/\sqrt{9}} = \frac{131.08 - 130}{1.5/\sqrt{9}} = 2.16$$

- ⑤ z does not fall in the rejection region. Therefore, H_0 cannot be rejected.

The data does not give strong evidence to claim that the true average differs from the design value of 130.

Test about a population mean with known σ

A manufacturer of sprinkler systems claims that the true average system-activation temperature is 130°F.

Using the same data as before, $n = 9$, sample average activation temperature of 131.08 °F, normal distribution and standard deviation 1.5 °F, significance level $\alpha = 0.05$

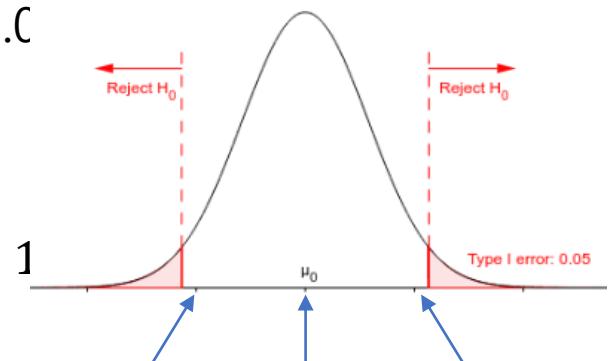
$$H_0 : \mu = 130 \quad \text{vs} \quad H_a : \mu \neq 130$$

$$\text{Rejection region : } \frac{\bar{x}-130}{1.5/\sqrt{9}} \leq -2.58 \rightarrow \bar{x} \leq 128.71, \quad \frac{\bar{x}-130}{1.5/\sqrt{9}} \geq 2.58 \rightarrow \bar{x} \geq 132.29$$

$$\begin{aligned}\beta(132) &= P(\text{type II error when } \mu= 132) \\ &= P(H_0 \text{ is not rejected when it is false because } \mu= 132) \\ &= P(128.71 \leq \bar{x} \leq 131.29 \text{ when } X \sim N(132, 1.5^2))\end{aligned}$$

$$131.29$$

$$= P\left(\frac{128.71-132}{1.5/\sqrt{9}} \leq Z \leq \frac{131.29-132}{1.5/\sqrt{9}}\right) = P(-6.58 \leq Z \leq -1.42) = 0.0778$$



$$128.71 \quad 130$$

β and sample size computation

Case I $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$ rejection region : $z \geq z_\alpha$ (upper tailed)

$$\begin{aligned}\beta(\mu') &= P(\text{type II error when } \mu=\mu') \\ &= P(H_0 \text{ is not rejected when } \mu=\mu') \\ &= P\left(Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \text{ when } \mu=\mu'\right) \\ &= P\left(\bar{X} < \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \text{ when } \mu=\mu'\right) \\ &= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu=\mu'\right) \\ &= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

β and sample size computation

Suppose we fix α and also specify β for an alternative value μ' .

Then we need,

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

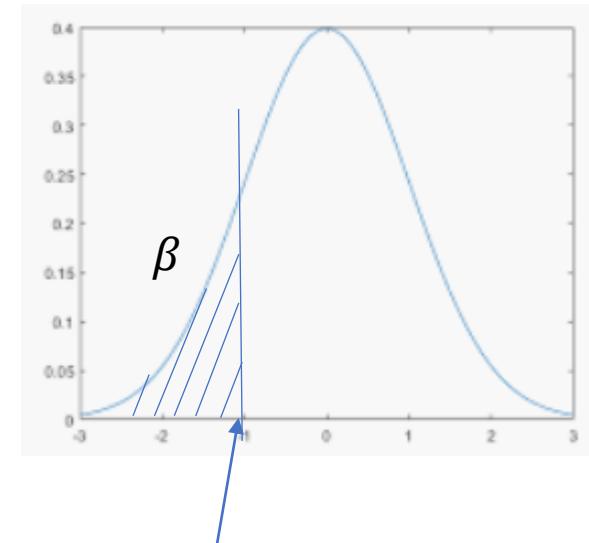
$$z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = -z_\beta \quad : z \text{ critical value}$$

that captures lower-tail area β

Therefore

$$\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = -(z_\alpha + z_\beta) \rightarrow \frac{\sigma}{\sqrt{n}} = \frac{\mu_0 - \mu'}{-(z_\alpha + z_\beta)}$$

$$\sqrt{n} = \frac{-\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \rightarrow n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \right]^2$$



$$z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \quad \text{and} \quad -z_\beta$$

β and sample size computation

Case I $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$ rejection region : $z \geq z_\alpha$ (upper tailed)

$$\begin{aligned}\beta(\mu') &= P(\text{type II error when } \mu=\mu') \\&= P(H_0 \text{ is not rejected when } \mu=\mu') \\&= P\left(Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \text{ when } \mu=\mu'\right) \\&= P\left(\bar{X} < \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \text{ when } \mu=\mu'\right) \\&= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu=\mu'\right) \\&= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

β and sample size computation

Suppose we fix α and also specify β for an alternative value μ' .

Then we need,

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

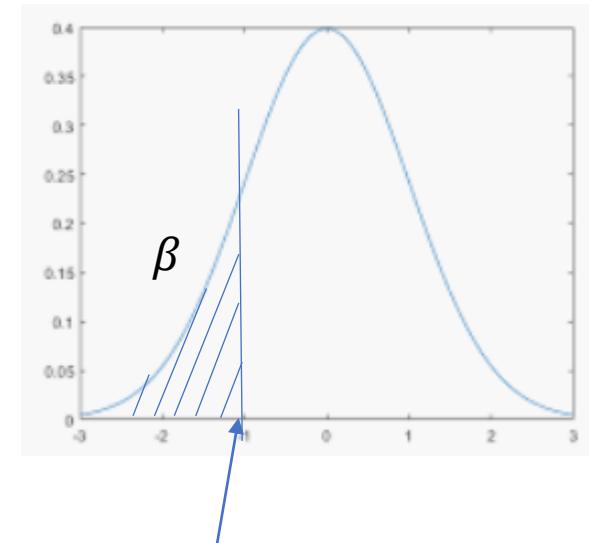
$$z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = -z_\beta \quad : z \text{ critical value}$$

that captures lower-tail area β

Therefore

$$\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = -(z_\alpha + z_\beta) \rightarrow \frac{\sigma}{\sqrt{n}} = \frac{\mu_0 - \mu'}{-(z_\alpha + z_\beta)}$$

$$\sqrt{n} = \frac{-\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \rightarrow n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \right]^2$$



$$z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \quad \text{and} \quad -z_\beta$$

β and sample size computation

Case II $H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$ rejection region : $z \leq -z_\alpha$ (lower tailed)

$$\begin{aligned}\beta(\mu') &= P(\text{type II error when } \mu=\mu') \\&= P(H_0 \text{ is not rejected when } \mu=\mu') \\&= P\left(Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha \text{ when } \mu=\mu'\right) \\&= P\left(\bar{X} > \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \text{ when } \mu=\mu'\right) \\&= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} > -z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu=\mu'\right) \\&= 1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

$P(Z < z) = \Phi(z)$ and $P(Z > z) = 1 - \Phi(z)$

β and sample size computation

Suppose we fix α and also specify β of type II error probability for an alternative value μ' .

Then we need,

$$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

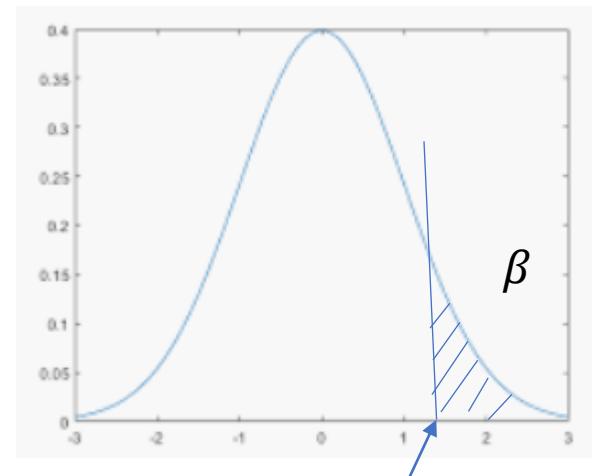
$$-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = z_\beta : z \text{ critical value}$$

that captures lower-tail area β

Therefore

$$\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = z_\alpha + z_\beta \rightarrow \frac{\sigma}{\sqrt{n}} = \frac{\mu_0 - \mu'}{z_\alpha + z_\beta}$$

$$\sqrt{n} = \frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \rightarrow n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \right]^2$$



$\beta(\mu')$ Summary

Alternative
Hypothesis

Type II Error Probability $\beta(\mu')$
for a Level α Test

$$H_a: \mu > \mu_0$$

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

$$H_a: \mu < \mu_0$$

$$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

$$H_a: \mu \neq \mu_0$$

$$\Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

where $\Phi(z)$ = the standard normal cdf.

The sample size n for which a level α test also has $\beta(\mu') = \beta$ at the alternative value μ' is

$$n = \begin{cases} \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \right]^2 & \text{for a one-tailed} \\ & \text{(upper or lower) test} \\ \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'} \right]^2 & \text{for a two-tailed test} \\ & \text{(an approximate solution)} \end{cases}$$

Example 8.7

Let μ denote the true average tread life of a certain type of tire.

Consider $H_0 : \mu = 30000$, vs $H_a : \mu > 30000$ based on a sample of size $n = 16$ from a normal population distribution with $\sigma = 1500$.

A test with $\alpha = 0.01$ requires $z_\alpha = z_{0.01} = 2.33$.

The probability of making a type II error when $\mu = 31000$ is,

$$\beta(31000) = \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \Phi\left(2.33 + \frac{30000 - 31000}{1500/\sqrt{16}}\right) = \Phi(-0.34) = 0.3669$$

When we want to make $\beta(31000) = 0.1 = \beta$, we have $z_\beta = z_{0.1} = 1.28$ and n becomes,

$$n = \left[\frac{1500(2.33 + 1.28)}{30000 - 31000} \right]^2 = [-5.42]^2 = 29.32$$

The sample size must be an integer, so $n = 30$ tires should be used.

Example 8.7

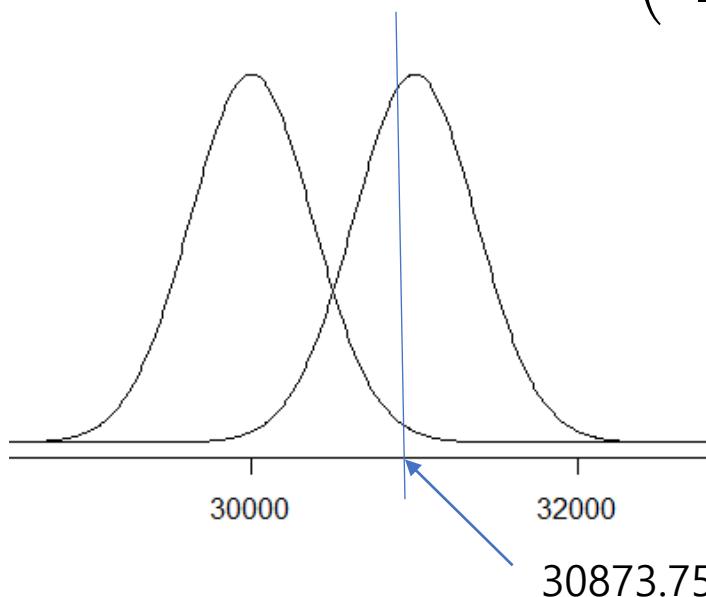
$H_0 : \mu = 30000$, vs $H_a : \mu > 30000$

Rejection region :

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - 30000}{1500/\sqrt{16}} > 2.33 = z_{0.01} \rightarrow \bar{X} > 30000 + 2.33 \times \frac{1500}{\sqrt{16}} = 30873.75$$

Type II error probability when $\mu = 31000$

$$\begin{aligned} P(\bar{X} < 30873.75 \text{ when } \mu = 31000) &= P\left(\frac{\bar{X} - 31000}{\frac{1500}{\sqrt{16}}} < \frac{30873.75 - 31000}{\frac{1500}{\sqrt{16}}}\right) \\ &= P(Z < -0.3366) = 0.3682 \end{aligned}$$



Test about a population mean with unknown σ and Large Sample

When $H_0 : \mu = \mu_0$ is true and n is large enough,

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0, 1)$$

$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ due to the central limit theorem

where s is the sample standard deviation.

Alternative Hypothesis

$$H_a : \mu > \mu_0$$

Rejection Region for level α test

$$z \geq z_\alpha \text{ (upper tailed)}$$

$$H_a : \mu < \mu_0$$

$$z \leq -z_\alpha \text{ (lower tailed)}$$

$$H_a : \mu \neq \mu_0$$

$$\text{either } z \geq z_{\frac{\alpha}{2}} \text{ or } z \leq -z_{\frac{\alpha}{2}} \text{ (two-tailed)}$$

Example 8.8

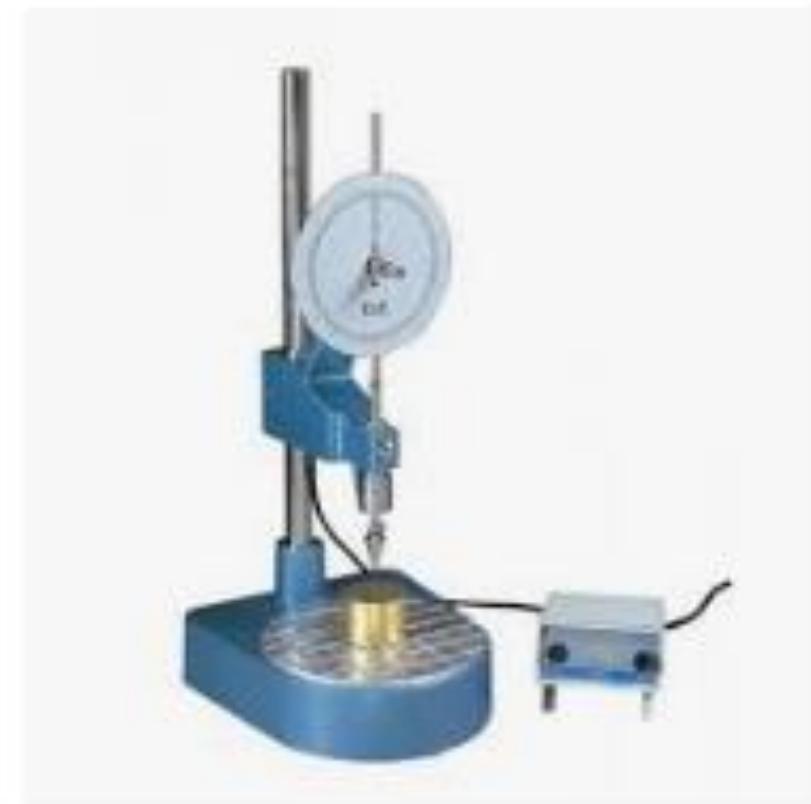
A dynamic cone penetrometer(DCP) is used for measuring material resistance to penetration.

It is required that true average DCP value for a certain type of pavement be less than 30.

(use $\alpha = 0.05$)

n	\bar{x}	s
52	28.7615	12.2647

Is the true average DCP value for a certain type of pavement be less than 30?



Example 8.8

- ① μ : true average DCP value
- ② $H_0 : \mu = 30$ versus $H_a : \mu < 30$
- ③ $z = \frac{\bar{x}-30}{s/\sqrt{n}}$
- ④ H_0 is rejected if $z \leq -z_{0.05} = -1.645$.
- ⑤ With $n = 52$, $\bar{x} = 28.76$, and $s = 12.2647$.

$$z = \frac{28.76-30}{12.2647/\sqrt{52}} = -0.73$$

- ⑥ Since $-0.73 > -1.645$, H_0 cannot be rejected.

Test about a population mean with unknown σ and Small Sample

- Suppose that X follows the normal distribution.
- When $H_0 : \mu = \mu_0$ is true,

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}. \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Test statistic value : $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for level α test
$H_a : \mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper tailed)
$H_a : \mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower tailed)
$H_a : \mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

Example 8.9

Glycerol contributes to the sweetness of wines.

The following observations shows the glycerol concentration for samples of wines :

2.67, 4.62, 4.14, 3.81, 3.83.

Suppose the desired concentration value is 4. Does the sample data suggest that true average concentration is something other than the desired value (use $\alpha = 0.05$)?

① $H_0 : \mu = 4$ versus $H_a : \mu \neq 4$

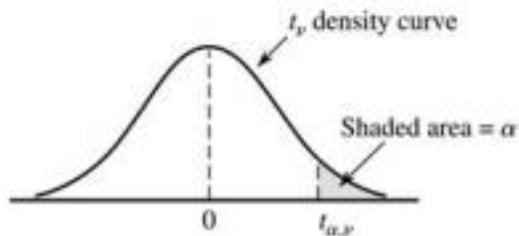
② $t = \frac{\bar{x} - 4}{s/\sqrt{n}}$

③ H_0 will be rejected if either $t \geq t_{\alpha/2, n-1} = t_{0.025, 4} = 2.776$ or $t \leq -2.776$:

④ Substituting $n = 5$, $\bar{x} = 3.814$, and $s = 0.718$, we get

$$t = \frac{3.814 - 4}{0.718/\sqrt{5}} = -0.58$$

⑤ Since $t = -0.58$ does not fall in the rejection region, H_0 cannot be rejected.

Table A.5 Critical Values for t Distributions

v	α						
	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781

```
> glycerol <- c(2.67, 4.62, 4.14, 3.81, 3.83)
> t.test(glycerol, mu=4)

One Sample t-test

data: glycerol
t = -0.57886, df = 4, p-value = 0.5937
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 2.921875 4.706125
sample estimates:
mean of x
 3.814
```

$$\text{P-value} = P(t_4 < -0.57886) + P(t_4 > 0.57886) = 0.5937$$

```
> pt(-0.57886,4) + 1 - (pt(0.57886,4)) #0.5937
```

Test on a proportion : Large-Sample Tests

Null hypothesis : $H_0 : p = p_0$

Test statistic value : $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ $\hat{p} \sim N(p_0, \frac{p_0(1-p_0)}{n})$

Alternative Hypothesis Rejection Region

$H_a : p > p_0$ $z \geq z_\alpha$ (upper tailed)

$H_a : p < p_0$ $z \leq -z_\alpha$ (lower tailed)

$H_a : p \neq p_0$ either $z \geq \frac{z_\alpha}{2}$ or $z \leq -\frac{z_\alpha}{2}$ (two-tailed)

These test procedures are valid provided that $np_0 > 10$, and $n(1 - p_0) > 10$.

Example 8.11

Natural cork in wine bottles is subject to deterioration.

Suppose that 16 of 91 bottles were considered spoiled by cork-associated characteristics.

Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? (use $\alpha = 0.1$)

(Solution) $H_0 : p=0.15$ vs $H_a : p>0.15$

Since $np_0 = 91(0.15) = 13.65 > 10$, and $n(1 - p_0) = 91(0.85) = 77.35 > 10$, the large sample z test can be used.

$$z = \frac{\hat{p} - 0.15}{\sqrt{(0.15)(0.85)/n}} = \frac{0.1758 - 0.15}{\sqrt{(0.15)(0.85)/91}} = 0.69$$

Since $z = 0.69 < z_{0.10} = 1.28$, H_0 cannot be rejected.

β and sample size computation

Case I $H_0 : p = p_0$ vs $H_a : p > p_0$ rejection region : $z \geq z_\alpha$ (upper tailed)

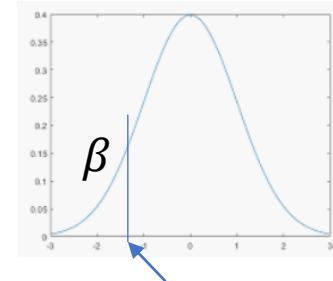
$$\begin{aligned}\beta(p') &= P(\text{type II error when } p=p') \\&= P(H_0 \text{ is not rejected when } p=p') \\&= P\left(Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < z_\alpha \text{ when } p=p'\right) \\&= P\left(\hat{p} < p_0 + z_\alpha \cdot \sqrt{p_0(1-p_0)/n} \text{ when } p=p'\right) \quad \hat{p} \sim N(p', \frac{p'(1-p')}{n}) \\&= P\left(\frac{\hat{p} - p'}{\sqrt{p'(1-p')/n}} < \frac{p_0 - p' + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} \text{ when } p=p'\right) \\&= \Phi\left(\frac{p_0 - p' + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right)\end{aligned}$$

β and sample size computation

Suppose we fix α and also specify β for an alternative value p' .

Then we need,

$$\Phi\left(\frac{p_0 - p' + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right) = \beta$$



This implies that

$$\frac{p_0 - p' + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} \text{ and } -z_\beta$$

$$\frac{p_0 - p' + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}} = -z_\beta \quad : z \text{ critical value that captures lower-tail area } \beta$$

Therefore

$$p_0 - p' = - \left(z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} + z_\beta \sqrt{\frac{p'(1-p')}{n}} \right) \rightarrow n = \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p'(1-p')}}{p' - p_0} \right]^2$$

$$(p_0 - p')^2 = \{z_\alpha^2(p_0(1-p_0)) + z_\beta^2(p'(1-p')) + z_\alpha z_\beta \sqrt{p_0(1-p_0)} \sqrt{p'(1-p')}}/n$$

$\beta(p')$ Summary

Alternative Hypothesis	$\beta(p')$
$H_a: p > p_0$	$\Phi\left[\frac{p_0 - p' + z_\alpha \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$
$H_a: p < p_0$	$1 - \Phi\left[\frac{p_0 - p' - z_\alpha \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$
$H_a: p \neq p_0$	$\Phi\left[\frac{p_0 - p' + z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right] - \Phi\left[\frac{p_0 - p' - z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$

The sample size n for which the level α test also satisfies $\beta(p') = \beta$ is

$$n = \begin{cases} \left[\frac{z_\alpha \sqrt{p_0(1 - p_0)} + z_\beta \sqrt{p'(1 - p')}}{p' - p_0} \right]^2 & \text{one-tailed test} \\ \left[\frac{z_{\alpha/2} \sqrt{p_0(1 - p_0)} + z_\beta \sqrt{p'(1 - p')}}{p' - p_0} \right]^2 & \text{two-tailed test (an approximate solution)} \end{cases}$$

P-Values

- Definition

The P-value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$H_0 : \mu = \mu_0 \quad \text{P-value}$$

$$H_a : \mu > \mu_0 \quad P(Z > z \text{ when } \mu = \mu_0) \quad z : \text{the value calculated from the available sample}$$

$$H_a : \mu < \mu_0 \quad P(Z < z \text{ when } \mu = \mu_0)$$

$$H_a : \mu \neq \mu_0 \quad P(Z < -|z| \text{ when } \mu = \mu_0) + P(Z > |z| \text{ when } \mu = \mu_0)$$

- Key points of the definition
 - The P-value is a probability.
 - This probability is calculated assuming that H_0 is true.

P-Values : Example 8.14

Urban storm water can be contaminated by many sources, including discarded batteries.

When ruptured, these batteries release metals of environmental significance.

A sample of 51 Panasonic AAA batteries gave a sample mean zinc mass of 2.06g and a sample standard deviation of 0.141g.

Does this data provide evidence for concluding that the population mean zinc mass exceeds 2.0g?

(Sol) $H_0 : \mu = 2.0$ versus $H_a : \mu > 2.0$.

The sample size is large enough to use CLT(Central Limit Theorem). The test statistic value is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.06 - 2.0}{0.141/\sqrt{51}} = 3.04$$

The P-value is

$$P(Z \geq 3.04 \text{ when } \mu = 2.0) = 1 - P(Z < 3.04 \text{ when } \mu = 2.0) = 0.0012$$

P-Values : Example 8.14

$H_0 : \mu = 2.0$ versus $H_a : \mu > 2.0$.

When the sample mean has a large value like 3, it will be a strong evidence that H_0 is false.

The larger the sample mean is, the stronger the evidence against H_0 becomes.

The observed sample mean is 2.06g.

When the sample mean is larger than 2.06g, it is more contradictory to null hypothesis than the currently observed mean 2.06g.

Therefore, the p-value is the probability that the sample mean is larger than 2.06g.

$$\text{P-value} = P(\bar{X} > 2.06) = P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > \frac{2.06 - 2.0}{0.141/\sqrt{51}}\right) = P(Z \geq 3.04)$$

More generally, the smaller the P-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.

That is, H_0 should be rejected in favor of H_a when the P-value is sufficiently small.

Decision rule based on the P-value

Select a significance level α . Then

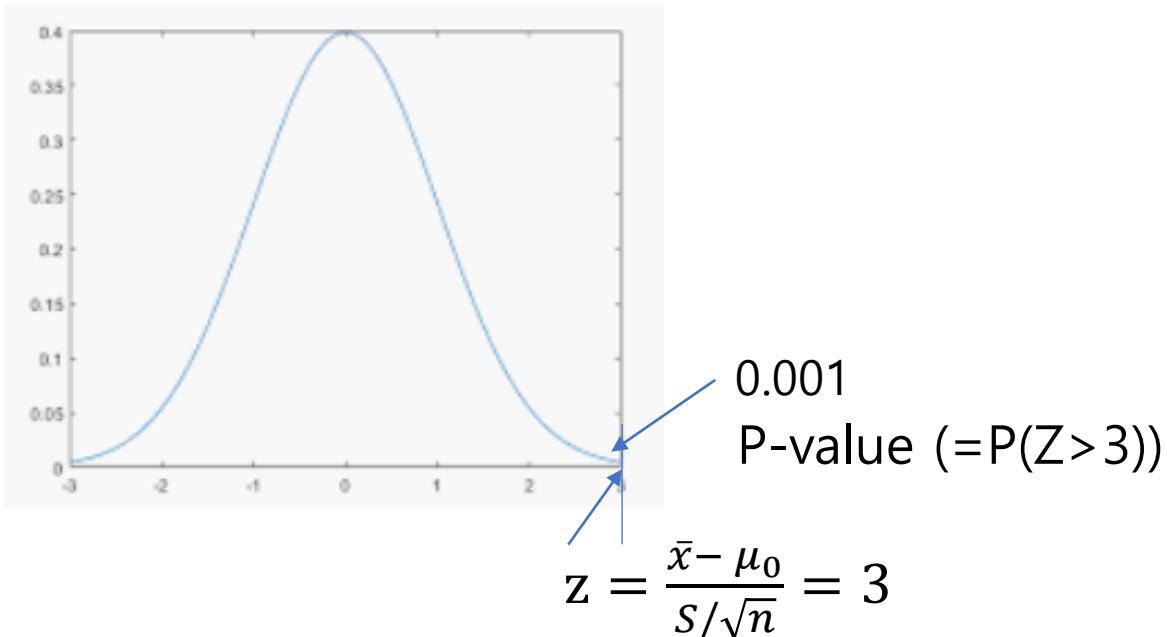
reject H_0 if P-value $\leq \alpha$

do not reject H_0 if P-value $> \alpha$

When the P-value is small like 0.001, it is very unlikely to observe the sample mean value under the H_0 . It is very unlikely to observe $z=3$ in standard normal distribution.

It means that H_0 is false.

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu > \mu_0.$$



- Thus if the P-value exceeds the chosen significance level, the null hypothesis cannot be rejected at that level.
- But if the P-value is equal to or less than the chosen significance level, then there is enough evidence to justify rejecting H_0 .
- In Example 14, we calculated P-value = 0.0012.
 - Then using a significance level of 0.01, we would reject the null hypothesis in favor of the alternative hypothesis because $0.0012 < 0.01$:
 - However, suppose we select a significance level of only 0.001, which requires more substantial evidence from the data before H_0 can be rejected.
 - In this case we would not reject H_0 because $0.0012 > 0.001$:

Example 8.12 : Computation of β

A package-delivery service advertises that at least 90% of all packages brought to its office by 9 A.M. for delivery in the same city are delivered by noon that day.

Consider $H_0 : p=0.9$ versus $H_a : p < 0.9$.

What is the type II error probability when $p=0.8$, if we use level $\alpha = 0.01$ test based on $n = 225$ packages?

What should the sample size be to ensure that $\beta(0.8) = 0.01$?

$$\begin{aligned}\beta(0.8) &= 1 - \Phi\left(\frac{p_0 - p' - z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right) \\ &= 1 - \Phi\left(\frac{0.9 - 0.8 - 2.33 \sqrt{\frac{0.9(1-0.9)}{225}}}{\sqrt{\frac{0.8(1-0.8)}{225}}}\right) = 0.0228 \\ n &= \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p'(1-p')}}{p' - p_0} \right]^2 = \left[\frac{2.33 \sqrt{0.9(1-0.9)} + 2.33 \sqrt{0.8(1-0.8)}}{0.8 - 0.9} \right]^2 = 266\end{aligned}$$

$\beta(p')$ Summary

Alternative Hypothesis	$\beta(p')$
$H_a: p > p_0$	$\Phi\left[\frac{p_0 - p' + z_\alpha \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$
$H_a: p < p_0$	$1 - \Phi\left[\frac{p_0 - p' - z_\alpha \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$
$H_a: p \neq p_0$	$\Phi\left[\frac{p_0 - p' + z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$ $- \Phi\left[\frac{p_0 - p' - z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p'(1 - p')/n}}\right]$

The sample size n for which the level α test also satisfies $\beta(p') = \beta$ is

$$n = \begin{cases} \left[\frac{z_\alpha \sqrt{p_0(1 - p_0)} + z_\beta \sqrt{p'(1 - p')}}{p' - p_0} \right]^2 & \text{one-tailed test} \\ \left[\frac{z_{\alpha/2} \sqrt{p_0(1 - p_0)} + z_\beta \sqrt{p'(1 - p')}}{p' - p_0} \right]^2 & \text{two-tailed test (an approximate solution)} \end{cases}$$

Example 8.12 - another solution

Rejection region :

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\hat{p} - 0.9}{\sqrt{0.9(1-0.9)/225}} < -z_\alpha = -z_{0.01} = -2.33$$

$$\hat{p} < 0.9 - 2.33\sqrt{0.9(1-0.9)/225} = 0.8534$$

$$\beta(0.8) = P(\hat{p} > 0.8534 \text{ when } p = 0.8)$$

$$= P\left(\frac{\hat{p} - 0.8}{\sqrt{0.8(1-0.8)/225}} > \frac{0.8534 - 0.8}{\sqrt{0.8(1-0.8)/225}}\right)$$

$$= P(Z > 2.0025) = 0.0226$$

Example 8.9

Glycerol contributes to the sweetness of wines.

The following observations shows the glycerol concentration for samples of wines :

2.67, 4.62, 4.14, 3.81, 3.83.

Suppose the desired concentration value is 4. Does the sample data suggest that true average concentration is something other than the desired value (use $\alpha = 0.05$)?

① $H_0 : \mu = 4$ versus $H_a : \mu \neq 4$

② $t = \frac{\bar{x}-4}{s/\sqrt{n}}$

③ H_0 will be rejected if either $t \geq t_{\alpha/2, n-1} = t_{0.025, 4} = 2.776$ or $t \leq -2.776$:

④ Substituting $n = 5$, $\bar{x} = 3.814$, and $s = 0.718$, we get

$$t = \frac{3.814-4}{0.718/\sqrt{5}} = -0.58$$

⑤ Since $t = -0.58$ does not fall in the rejection region, H_0 cannot be rejected.

```
> glycerol <- c(2.67, 4.62, 4.14, 3.81, 3.83)
> t.test(glycerol, mu=4)

One Sample t-test

data: glycerol
t = -0.57886, df = 4, p-value = 0.5937
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 2.921875 4.706125
sample estimates:
mean of x
 3.814
```

$$\text{P-value} = P(t_4 < -0.57886) + P(t_4 > 0.57886) = 0.5937$$

```
> pt(-0.57886,4) + 1 - (pt(0.57886,4)) #0.5937
```

P-Values for t distribution

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

$$H_0 : \mu = \mu_0$$

P-value

$$H_a : \mu > \mu_0 \quad P(t_{n-1} > t \text{ when } \mu = \mu_0) \quad t : \text{the value calculated from the available sample}$$

$$H_a : \mu < \mu_0 \quad P(t_{n-1} < t \text{ when } \mu = \mu_0)$$

$$H_a : \mu \neq \mu_0 \quad P(t_{n-1} < -|t| \text{ when } \mu = \mu_0) + P(t_{n-1} > |t| \text{ when } \mu = \mu_0)$$

Hypothesis Test of mean

Data on the modulus of rupture of composite beams designed to add value to low-grade sweetgum lumber is given in an article in J. of Bridge Engr.

6807.99, 7637.06, 6663.28, 6165.03, 6991.41, 6992.23, 6981.46, 7569.75,

7437.88, 6872.39, 7663.18, 6032.28, 6906.04, 6617.17, 6984.12, 7093.71,

7659.50, 7378.61, 7295.54, 6702.76, 7440.17, 8053.26, 8284.75, 7347.95,

7422.69, 7886.87, 6316.67, 7713.65, 7503.33, 7674.99

Does this data provide evidence for concluding that the population mean modulus of rupture exceeds 7000 (use $\alpha = 0.05$)?

(Sol) $H_0 : \mu = 7000$ versus $H_a : \mu > 7000$.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7203.19 - 7000}{543.54/\sqrt{30}} = 2.0475$$

① $H_0 : \mu = 7000$ versus $H_a : \mu > 7000$.

② $t = \frac{\bar{x} - 7000}{s/\sqrt{n}}$

③ H_0 will be rejected if either $t \geq t_{\alpha, n-1} = t_{0.05, 29} = 1.6991$

④ Substituting $n = 30$, $\bar{x} = 7203.19$, and $s = 543.54$, we get

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7203.19 - 7000}{543.54/\sqrt{30}} = 2.0475$$

⑤ Since $t = 2.0475$ falls in the rejection region, H_0 can be rejected.

⑥ Since $p\text{-value} = P(t_{29} > 2.0475) = 0.0249 < 0.05$, H_0 can be rejected.

> 1-pt(2.0475, 29) # 0.0249

- modulus <- c(6807.99, 7637.06, 6663.28, 6165.03, 6991.41, 6992.23, 6981.46, 7569.75, 7437.88, 6872.39, 7663.18, 6032.28, 6906.04, 6617.17, 6984.12, 7093.71, 7659.50, 7378.61, 7295.54, 6702.76, 7440.17, 8053.26, 8284.75, 7347.95, 7422.69, 7886.87, 6316.67, 7713.65, 7503.33, 7574.00)
 > t.test(modulus, mu=7000, alternative="greater")

 One Sample t-test

 data: modulus
 t = 2.0475, df = 29, p-value = 0.02488
 alternative hypothesis: true mean is greater than 7000
 95 percent confidence interval:
 7034.575 Inf
 sample estimates:
 mean of x
 7203.191

Chapter 9 – Inference Based on Two Samples

Difference Between Two Population Means

- Basic Assumptions

1. X_1, X_2, \dots, X_m is a **random sample** from a distribution with mean μ_1 and variance σ_1^2 .
2. Y_1, Y_2, \dots, Y_n is a **random sample** from a distribution with mean μ_2 and variance σ_2^2 .
3. The X and Y samples are **independent** of one another.

- Proposition

The expected value of $\bar{X} - \bar{Y}$ is $\mu_1 - \mu_2$, so $\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$.

The standard deviation of $\bar{X} - \bar{Y}$ is

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

The sample variances must be used to estimate this when population variances are unknown.

$$E(\bar{X})=\mu_1, V(\bar{X})=\frac{\sigma_1^2}{m}, E(\bar{Y})=\mu_2, V(\bar{Y})=\frac{\sigma_2^2}{n} \rightarrow E(\bar{X} - \bar{Y})=\mu_1 - \mu_2, V(\bar{X} - \bar{Y})=\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

Test Procedures for Normal Populations with Known Variances

- If both of the population distributions are normal, the difference $\bar{X} - \bar{Y}$ is normally distributed, with expected value $\mu_1 - \mu_2$ and standard deviation $\sigma_{\bar{X}-\bar{Y}}$.
- Standardizing $\bar{X} - \bar{Y}$ gives the standard normal variable

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

Test about a population means with Known σ

Null hypothesis : $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic value :

$$Z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \quad \bar{X} - \bar{Y} \sim N \left(\Delta_0, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \right)$$

Alternative Hypothesis	Rejection Region for level α test
$H_a : \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (upper tailed)
$H_a : \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (lower tailed)
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\frac{\alpha}{2}}$ or $z \leq -z_{\alpha/2}$ (two- tailed)

Example 9.1

Analysis of a random sample consisting of $m = 20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\bar{x} = 29.8$ ksi.

A second random sample of $n = 25$ two-sided galvanized steel specimens gave a sample average strength of $\bar{y} = 34.7$ ksi.

Assuming that the two yield-strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, does the data indicate that the corresponding true average yield strengths μ_1 and μ_2 are different?

Let's carry out a test at significance level $\alpha = 0.01$.

- ① Parameter of interest: $\mu_1 - \mu_2$
- ② Null hypothesis: $H_0 : \mu_1 - \mu_2 = 0$
- ③ Alternative hypothesis: $H_a : \mu_1 - \mu_2 \neq 0$
- ④ With $\Delta_0 = 0$, the test statistic value is

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

Example 9.1

⑤ $\alpha = 0.01 \rightarrow H_0$ will be rejected if $z \geq 2.58 = z_{0.01/2}$ or $z \leq -2.58 = -z_{0.005}$

⑥ Test statistic value: $z = \frac{29.8 - 34.7}{\sqrt{\frac{16.0}{20} + \frac{25.0}{25}}} = -3.66$

⑥ Since $-3.66 < -2.58$, z falls in the lower tail of the rejection region.

Therefore, H_0 is rejected at level 0.01.

Large-Sample Tests

The assumptions of normal population distributions and known values of σ_1 and σ_2 are fortunately unnecessary when both sample sizes are sufficiently large.

WHY?

Furthermore, using S_1^2 and S_2^2 in place of σ_1^2 and σ_2^2 gives a variable whose distribution is approximately standard normal:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

These tests are usually appropriate if both $m > 40$ and $n > 40$.

Example 9.4

Data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds by more than 200 calories per day the true average intake for those who don't typically eat fast food?

Let's investigate by carrying out a test of hypotheses at a significance level of approximately 0.05.

Example 9.4

- ① Null hypothesis: $H_0 : \mu_1 - \mu_2 = -200$ vs $\mu_1 - \mu_2 < -200$

μ_1 : the true average calories intake for teens who don't typically eat fast food

- ② The test statistic value is

$$Z = \frac{\bar{x} - \bar{y} - (-200)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

- ③ $\alpha = 0.05 \rightarrow H_0$ will be rejected if $z \leq -z_{0.05} = -1.645$

- ④ Test statistic value: $z = \frac{2258 - 2637 + 200}{\sqrt{\frac{(1519)^2}{663} + \frac{(1138)^2}{413}}} = -2.20$

- ⑤ Since $-2.20 < -1.645$, H_0 is rejected at level 0.05.

- ⑥ The P -value for the test is

$$P\text{-value} = P(Z < -2.20) = \Phi(-2.20) = 0.0139$$

- ③ Because $0.0139 \leq 0.05$, we reject H_0 at significance level 0.05.

Confidence Intervals for $\mu_1 - \mu_2$

- Since the area under the z curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

- Manipulation of the inequalities inside the parentheses to isolate $\mu_1 - \mu_2$ yields

$$-z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \bar{X} - \bar{Y} - (\mu_1 - \mu_2) < z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

$$-(\bar{X} - \bar{Y}) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < -(\mu_1 - \mu_2) < -(\bar{X} - \bar{Y}) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

$$(\bar{X} - \bar{Y}) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < (\mu_1 - \mu_2) < (\bar{X} - \bar{Y}) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Confidence Intervals for $\mu_1 - \mu_2$

- We have equivalent probability statement

$$P\left(\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha$$

$$\bar{X} - \bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

- Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ with a confidence level of approximately $100(1 - \alpha)$ % is

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$\bar{x} - \bar{y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

Example 9.5

An experiment carried out to study various characteristics of anchor bolts resulted in 78 observations on shear strength (kip) of 3/8-in. diameter bolts and 88 observations on the strength of 1/2-in. diameter bolts. The summaries are as follows :

Variable	Sample Size	Sample Mean	Sample SD
diam 3/8	78	4.25	1.3
diam 1/2	88	7.14	1.68

Let's now calculate a confidence interval for the difference between true average shear strength for 3/8-in. bolts (μ_1) and true average shear strength for 1/2-in. bolts (μ_2) using a confidence level of 95%:

$$4.25 - 7.14 \pm (1.96) \sqrt{\frac{(1.30)^2}{78} + \frac{(1.68)^2}{88}} = -2.89 \pm (1.96)(0.2318) = (-3.34, -2.44)$$

The Two-Sample t Test and Confidence Interval

- When the population distribution are both normal, the standardized variable

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

has approximately a t distribution with df v estimated from the data by

$$v = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n} \right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

The Two-Sample t Test and Confidence Interval

- The **two-sample t confidence interval for $\mu_1 - \mu_2$** with confidence level $100(1 - \alpha)$ % is

$$P\left(-t_{\frac{\alpha}{2}, v} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} < t_{\frac{\alpha}{2}, v}\right) = 1 - \alpha$$

$$\bar{X} - \bar{Y} \pm t_{\frac{\alpha}{2}, v} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

- The **two-sample t test** for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is as follows:

Test statistic value: $t = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$

Alternative Hypothesis

$$H_a : \mu_1 - \mu_2 > \Delta_0$$

$$H_a : \mu_1 - \mu_2 < \Delta_0$$

$$H_a : \mu_1 - \mu_2 \neq \Delta_0$$

Rejection Region for level α test

$$t \geq t_{\alpha, v} \text{ (upper tailed)}$$

$$t \leq -t_{\alpha, v} \text{ (lower tailed)}$$

$$\text{either } t \geq t_{\alpha/2, v} \text{ or } t \leq -t_{\alpha/2, v} \text{ (two-tailed)}$$

Example 9.6

- Consider the following data on two different types of plain-weave fabric:

Fabric Type	Sample Size	Sample Mean	Sample SD	Sample Variance
Cotton	10	51.71	0.79	0.6241
Triacetate	10	136.14	3.59	12.8881

Assuming that the porosity distributions for both types of fabric are normal, let's calculate a confidence interval for the difference between true average porosity for the cotton fabric and that for the acetate fabric, using a 95% confidence level

$$v = \frac{\left(\frac{0.6241}{10} + \frac{12.8881}{10}\right)^2}{\frac{(0.6241/10)^2}{9} + \frac{(12.8881/10)^2}{9}} = 9.84$$

Thus we use $v = 9$.

$$\bar{x} - \bar{y} \pm t_{\frac{\alpha}{2}, v} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = 51.71 - 136.14 \pm 2.262 \sqrt{\frac{0.6241}{10} + \frac{12.8881}{10}} = (-87.06, -81.80)$$

Pooled t Procedures

- Two population distributions are normal and they have equal variances ($\sigma_1^2 = \sigma_2^2$)
- Pooled estimator of σ^2

$$s_p^2 = \frac{m-1}{m+n-2} s_1^2 + \frac{n-1}{m+n-2} s_2^2 = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{m+n-2}$$

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{m} + \frac{1}{n} \right)}} \sim N(0,1)$$

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{m} + \frac{1}{n} \right)}} \sim t_{m+n-2}$$

The Two-Sample t Test and Confidence Interval

- The **two-sample t confidence interval for $\mu_1 - \mu_2$** with confidence level $100(1 - \alpha)$ % is

$$\bar{X} - \bar{Y} \pm t_{\frac{\alpha}{2}, m+n-2} s_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

- The **two-sample t test** for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is as follows:

Test statistic value: $t = \frac{\bar{x} - \bar{y} - \Delta_0}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$

Alternative Hypothesis

$$H_a : \mu_1 - \mu_2 > \Delta_0$$

$$H_a : \mu_1 - \mu_2 < \Delta_0$$

$$H_a : \mu_1 - \mu_2 \neq \Delta_0 \text{ (two-tailed)}$$

Rejection Region for level α test

$$t \geq t_{\alpha, m+n-2} \text{ (upper tailed)}$$

$$t \leq -t_{\alpha, m+n-2} \text{ (lower tailed)}$$

$$\text{either } t \geq t_{\alpha/2, m+n-2} \text{ or } t \leq -t_{\alpha/2, m+n-2} \text{ (two-tailed)}$$

Example (<https://online.stat.psu.edu/stat414/node/201/>)

The feeding habits of two species of net-casting spiders are studied. The species, the deinopis and menneus, coexist in eastern Australia. The following data were obtained on the size, in millimeters, of the prey of

random samples of the two species:

deinopis : 12.9, 10.2, 7.4, 7.0, 10.5, 11.9, 7.1, 9.9, 14.4, 11.3

menneus : 10.2, 6.9, 10.9, 11.0, 10.1, 5.3, 7.5, 10.3, 9.2, 8.8

Compute 95% confident interval for the actual mean difference in the size of the prey .

$$s_{\text{deinopis}}^2 = 6.318, s_{\text{menneus}}^2 = 3.597, \bar{x} = 10.26, \bar{Y} = 9.02$$

$$s_p^2 = \frac{10-1}{10+10-2} 6.318 + \frac{10-1}{10+10-2} 3.597 = 4.9575$$

95% confident interval

$$\bar{X} - \bar{Y} \pm t_{\frac{\alpha}{2}, m+n-2} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 10.26 - 9.02 \pm (2.101) \sqrt{4.9575} \sqrt{\frac{1}{10} + \frac{1}{10}} = 1.24 \pm 2.092 \text{ or } (-0.852, 3.332)$$

```
> deinopis <- c(12.9, 10.2, 7.4, 7.0, 10.5, 11.9, 7.1, 9.9, 14.4, 11.3)
> menneus <- c(10.2, 6.9, 10.9, 11.0, 10.1, 5.3, 7.5, 10.3, 9.2, 8.8)
> t.test(deinopis, menneus, var.equal=T)
```

Two sample t-test

```
data: deinopis and menneus
t = 1.2453, df = 18, p-value = 0.229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

-0.8520327 3.3320327

sample estimates:

mean of x mean of y
10.26 9.02

P-value : $P(t_{18} < -1.2453) + P(t_{18} > 1.2453) = 0.229$

```
> pt(-1.2453, 18)+(1-pt(1.2453, 18)) # 0.229
```

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{10.26 - 9.02}{\sqrt{4.9575} \sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.2453$$

Suppose the variances are different

- $\bar{x} = 10.26, \bar{y} = 9.02, s_{\text{deinopis}}^2 = 6.318, s_{\text{menneus}}^2 = 3.597$

$$v = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = \frac{\left(\frac{6.318}{10} + \frac{3.597}{10}\right)^2}{\frac{(6.318/10)^2}{9} + \frac{(3.597/10)^2}{9}} = 16.74$$

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{10.26 - 9.02}{\sqrt{\frac{6.318}{10} + \frac{3.597}{10}}} = 1.2453$$

- 95% CI of $\mu_1 - \mu_2$: use $v = 16$ (largest integer less than 16.74)

$$\bar{x} - \bar{y} \pm t_{\frac{\alpha}{2}, 16} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = 10.26 - 9.02 \pm 2.12 \sqrt{\frac{6.318}{10} + \frac{3.597}{10}} = (-0.871, 3.351)$$

$$\bar{x} - \bar{y} \pm t_{\frac{\alpha}{2}, 16.74} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = 10.26 - 9.02 \pm 2.112 \sqrt{\frac{6.318}{10} + \frac{3.597}{10}} = (-0.863, 3.343)$$

```
> deinopis <- c(12.9, 10.2, 7.4, 7.0, 10.5, 11.9, 7.1, 9.9, 14.4, 11.3)
> menneus <- c(10.2, 6.9, 10.9, 11.0, 10.1, 5.3, 7.5, 10.3, 9.2, 8.8)
> t.test(deinopis, menneus)
```

Welch Two Sample t-test

```
data: deinopis and menneus
t = 1.2453, df = 16.74, p-value = 0.2302
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.8633815 3.3433815
sample estimates:
mean of x mean of y
10.26      9.02
```

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_a : \mu_1 - \mu_2 \neq 0$$

$$\text{P-value} : P(t_{16.74} < -1.2453) + P(t_{16.74} > 1.2453) = 0.2302$$

```
> pt(-1.2453, 16.74)+(1-pt(1.2453, 16.74)) # 0.2302
```

Analysis of Paired Data

- There are a number of experimental situations in which there is only one set of n individuals or experimental objects: making two observations on each one results in a natural pairing of values.
- Assumption

The data consists of n independently selected pairs

$(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$, with $E(X_i) = \mu_1$, and $E(Y_i) = \mu_2$.

Let $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \dots D_n = X_n - Y_n$ so the D_i 's are the differences within pairs.

Then the D_i 's are assumed to be normally distributed with mean value μ_D and variance σ_D^2 .
$$D_i \sim N(\mu_D, \sigma_D^2)$$

Example

- Suppose that you want to test the effect of a diet plan.
- A diet plan may consist of controlling the food and making some exercise.
- So you select 10 volunteers.
- You measure their weight before starting a diet plan.
- Make them go through the diet plan for 3 months.
- You measure their weight again.
- You will test if the diet plan actually lessens their weight.
- Test of tread wear on tires
- A manufacturer wishes to compare the tread wear of tires made of a new material with that of tires made of a conventional material.
- One tire of each type is placed on each front wheel of 10 front-wheel-drive automobiles.
- Each car is driven for 40,000, a measurement of tread wear is then made on each tire.

Analysis of Paired Data

Null hypothesis : $H_0 : \mu_D = \Delta_0$

Test statistic value :

$$t = \frac{\bar{d} - \Delta_0}{s_D / \sqrt{n}}$$

Alternative Hypothesis	Rejection Region for level α test
$H_a : \mu_D > \Delta_0$	$t \geq t_{\alpha, n-1}$ (upper tailed)
$H_a : \mu_D < \Delta_0$	$t \leq -t_{\alpha, n-1}$ (lower tailed)
$H_a : \mu_D \neq \Delta_0$	either $t \geq t_{\alpha/2, n-1}$ or $t \leq -t_{\alpha/2, n-1}$ (two-tailed)

100(1- α)% confidence interval of μ :

$$\bar{d} \pm t_{\alpha/2, n-1} \frac{s_D}{\sqrt{n}}$$

Example 9.9

The accompanying data was obtained from a sample of n=16 subjects.

Each observation is the amount of time, expressed as a proportion of total time observed, during which arm elevation was below 30°.

The two measurements from each subject were obtained 18 months apart. During this period, work conditions were changed.

Subject	1	2	3	4	5
6	81	87	86	82	90
Before					
86	96	73			
After					
67	92	70	78	91	78
Difference			3	-4	8
19	4	3		4	6
Subject	9	10	11	12	13
14	71	77	73	88	88
15					
16					

Example 9.9

① $H_0 : \mu_D = 0$ vs $H_a : \mu_D \neq 0$

② Test statistic : $t = \frac{\bar{d} - \Delta_0}{s_D / \sqrt{n}}$

③ Test statistic value: $t = \frac{6.75}{8.234 / \sqrt{16}} = 3.28$

④ The P -value for the test is

$$P\text{-value} = 2 \times P(T_{15} > 3.28) = 0.00506$$

⑤ Because $0.00506 \leq 0.01$, we can reject H_0 at significance level 0.05 or 0.01.

⑥ Since $t = 3.28 > t_{\frac{0.01}{2}, 15} = 2.947$, H_0 can be rejected at significance level 0.01.

$> 2 * (1 - pt(3.28, 15))$

[1] 0.00506228

```
> before <- c(81, 87, 86, 82, 90, 86, 96, 73, 74, 75, 72, 80, 66, 72, 56, 82)
> after <- c(78, 91, 78, 78, 84, 67, 92, 70, 58, 62, 70, 58, 66, 60, 65, 73)
> diff <- before - after
> t.test(diff)
```

One Sample t-test

```
data: diff
t = 3.2791, df = 15, p-value = 0.005072
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.362371 11.137629
sample estimates:
mean of x
 6.75
```

$$\bar{d} \pm t_{\alpha/2, n-1} \frac{s_D}{\sqrt{n}} = 6.75 \pm 2.131 \frac{8.234}{\sqrt{16}} \rightarrow (2.362, 11.138)$$

Inference Concerning a Difference Between Population Proportions

- Let

p_1 = the true proportion of *Ss* in population # 1

p_2 = the true proportion of *Ss* in population # 2

m : size of a sample selected from the first population

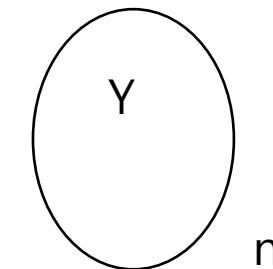
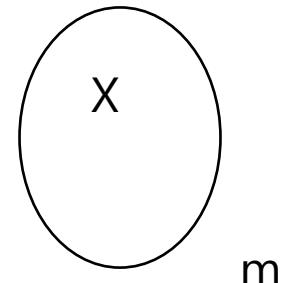
n : size of a sample selected from the second population

X : the number of *Ss* in the first sample

Y : the number of *Ss* in the second sample

$$\hat{p}_1 = \frac{X}{m}$$

$$\hat{p}_2 = \frac{Y}{n}$$



Inference Concerning a Difference Between Population Proportions

- Proposition

Let $\hat{p}_1 = X/m$ and $\hat{p}_2 = Y/n$ where $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$ with X and Y independent variables.

$$E(\hat{p}_1) = p_1, V(\hat{p}_1) = \frac{p_1 q_1}{m}, E(\hat{p}_2) = p_2, V(\hat{p}_2) = \frac{p_2 q_2}{n},$$

Then $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

So $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$, and

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \quad (\text{where } q_i = 1 - p_i)$$

A Large-Sample Test Procedures

The most general null hypothesis an investigator might consider would be of the form

$$H_0: p_1 - p_2 = \Delta_0.$$

Since the vast majority of actual problems of this sort involve $\Delta_0 = 0$ (i.e., the null hypothesis $H_0: p_1 = p_2$). we'll concentrate on this case.

When $H_0: p_1 - p_2 = 0$ is true, let p denote the common value of p_1 and p_2 (and similarly for q).

Then the standardized variable $(p_1 = p_2 = p)$

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{pq}{m} + \frac{pq}{n}}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

has approximately a standard normal distribution when H_0 is true.

The natural estimator of p is then

$$\hat{p} = \frac{X+Y}{m+n} = \frac{m}{m+n} \cdot \hat{p}_1 + \frac{n}{m+n} \cdot \hat{p}_2$$

A Large-Sample Test Procedures

Null hypothesis : $H_0 : p_1 - p_2 = 0$

Test statistic value : $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$

Alternative Hypothesis

$H_a : p_1 - p_2 > 0$

$H_a : p_1 - p_2 < 0$

$H_a : p_1 - p_2 \neq 0$

Rejection Region for level α test

$z \geq z_\alpha$ (upper tailed)

$z \leq -z_\alpha$ (lower tailed)

either $z \geq z_{\frac{\alpha}{2}}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

The test can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$, and $n\hat{q}_2$ are all at least 10.

Example 9.11

An article reported that of 549 participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths.

Does this data suggest that the regular use of aspirin after diagnosis will decrease the incidence rate of colorectal cancer-specific deaths?

Let's test the appropriate hypotheses using a significance level of 0.05.

$$H_0 : p_1 - p_2 = 0 \quad \text{vs} \quad H_a : p_1 - p_2 < 0$$

$$\hat{p}_1 = \frac{81}{549} = 0.1475, \hat{p}_2 = \frac{141}{730} = 0.1932, \hat{p} = \frac{81+141}{549+730} = 0.1736$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{0.1475 - 0.1932}{\sqrt{(0.1736)(0.8264)\left(\frac{1}{549} + \frac{1}{730}\right)}} = -2.14$$

$$\text{P-value} : P(Z < -2.14) = 0.0162.$$

Because $0.0162 \leq 0.05$, the null hypothesis can be rejected at significance level 0.05.

Or $Z = -2.14 < -z_{0.05} = -1.645$, z falls in the rejection region, and H_0 can be rejected at $\alpha = 0.05$.

A Large-Sample Confidence Interval

A CI for $p_1 - p_2$ with confidence level approximately $100(1 - \alpha)\%$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}{n}}$$

This interval can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$, and $n\hat{q}_2$ are all at least 10.

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}}}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$-z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}} < \hat{p}_1 - \hat{p}_2 - (p_1 - p_2) < z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}}$$

$$-(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}} < -(p_1 - p_2) < -(\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\hat{p} \hat{q} \left(\frac{1}{m} + \frac{1}{n}\right)}$$

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}}$$

Example 9.12

Comparison of the cancer treatment method :

Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long.

What is the 99% confidence interval for this difference in proportions?

$$\hat{p}_1 = \frac{76}{154} = 0.494, \quad \hat{p}_2 = \frac{98}{164} = 0.598$$

$$\begin{aligned}\hat{p}_1 - \hat{p}_2 &\pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}{n}} = 0.494 - 0.598 \pm 2.58 \sqrt{\frac{(0.494)(0.506)}{154} + \frac{(0.598)(0.402)}{164}} \\ &= (-0.247, 0.039)\end{aligned}$$

Example 9.12

Comparison of the cancer treatment method :

Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long.

What is the 99% confidence interval for this difference in proportions?

$$\hat{p}_1 = \frac{76}{154} = 0.494, \quad \hat{p}_2 = \frac{98}{164} = 0.598$$

$$\begin{aligned}\hat{p}_1 - \hat{p}_2 &\pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{m} + \frac{\hat{p}_2 \hat{q}_2}{n}} = 0.494 - 0.598 \pm 2.58 \sqrt{\frac{(0.494)(0.506)}{154} + \frac{(0.598)(0.402)}{164}} \\ &= (-0.247, 0.039)\end{aligned}$$

Inference Concerning Two Population Variances

- $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_a : \sigma_1^2 > \sigma_2^2$
- The F Distribution

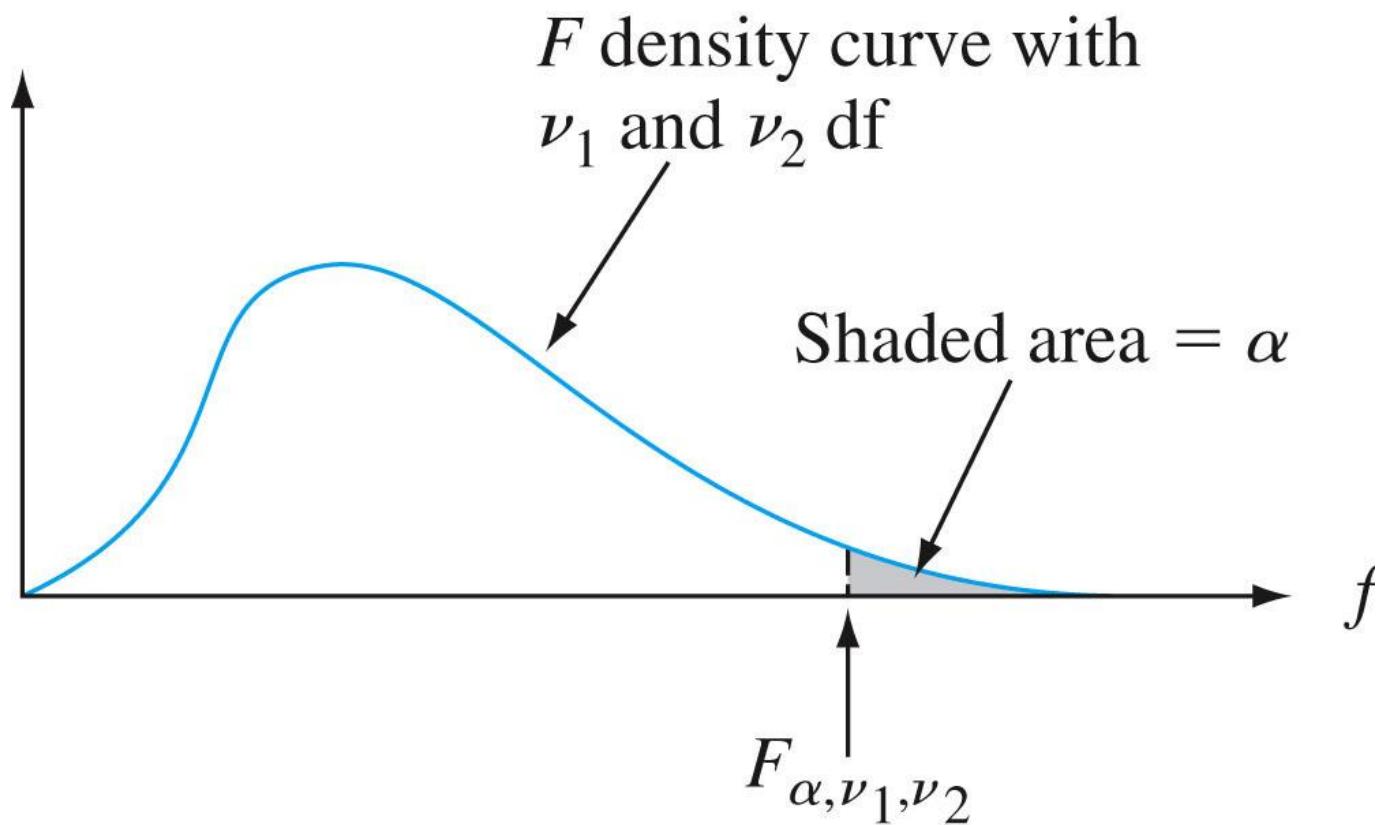
The F probability distribution has two parameters, denoted by ν_1 and ν_2 .

The parameter ν_1 is called the *numerator degrees of freedom*, and ν_2 is the *denominator degrees of freedom*.

If X_1 and X_2 are independent chi-squared rv's with ν_1 and ν_2 df, respectively, then the rv

$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

(the ratio of the two chi-squared variables divided by their respective degrees of freedom), can be shown to have an F distribution.



© 2007 Thomson Higher Education

Fig. 9-8, p. 360

The F Test for Equality of Variances

- Let X_1, \dots, X_m be a random sample from a normal distribution with variance σ_1^2 ,
Let Y_1, \dots, Y_n be another random sample (independent of the X_i 's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances.
- Then the rv

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

- This theorem results from combining the fact that the variables $(m - 1)s_1^2/\sigma_1^2$ and $(n - 1)s_2^2/\sigma_2^2$ each have a chi-squared distribution with $m - 1$ and $n - 1$ df, respectively.
- The claim that $\sigma_1^2 = \sigma_2^2$ is then rejected if the ratio differs by too much from 1 .

The F Test for Equality of Variances

Let X_1, \dots, X_m be a random sample from a normal distribution with mean μ_1 and variance σ_1^2

$$\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_1}{\sigma_1}\right)^2 + \dots + \left(\frac{x_m - \mu_1}{\sigma_1}\right)^2 = \frac{\sum_{i=1}^m (x_i - \mu_1)^2}{\sigma_1^2} \sim \chi_m^2$$

$$\left(\frac{x_1 - \bar{x}_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \bar{x}_1}{\sigma_1}\right)^2 + \dots + \left(\frac{x_m - \bar{x}_1}{\sigma_1}\right)^2 = \frac{\sum_{i=1}^m (x_i - \bar{x}_1)^2}{\sigma_1^2} = \frac{(m-1)s_1^2}{\sigma_1^2} \sim \chi_{m-1}^2$$

$$s_1^2 = \frac{\sum_{i=1}^m (x_i - \bar{x}_1)^2}{m-1}$$

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} = \frac{\frac{(m-1)s_1^2}{\sigma_1^2}/(m-1)}{\frac{(n-1)s_2^2}{\sigma_2^2}/(n-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{m-1, n-1}$$

The F Test for Equality of Variances

Null hypothesis : $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value : $f = \frac{s_1^2}{s_2^2}$

Alternative Hypothesis Rejection Region for level α test

$H_a : \sigma_1^2 > \sigma_2^2$ $f \geq F_{\alpha, m-1, n-1}$ (upper tailed)

$H_a : \sigma_1^2 < \sigma_2^2$ $f \leq F_{1-\alpha, m-1, n-1}$ (lower tailed)

$H_a : \sigma_1^2 \neq \sigma_2^2$ either $f \geq F_{\alpha/2, m-1, n-1}$ or $f \leq F_{1-\alpha/2, m-1, n-1}$ (two-tailed)

Example 9.14

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$.

For 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion that the ferritin distribution in the elderly had a smaller variance than in the younger adults? Use $\alpha = 0.01$.

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs } H_1 : \sigma_1^2 < \sigma_2^2$$

$$f = \frac{52.6^2}{84.2^2} = 0.390$$

Since $f = 0.390 < F_{0.99,27,25} = 0.394$, H_0 is rejected at level 0.01.

```
> qf(0.01, 27, 25)
```

```
[1] 0.3943183
```

Chapter 12 – Simple Linear Regression and Correlation

Outline

- ① Simple Linear Regression
- ② Estimating parameters
- ③ Inferences about the slope parameter β_1
- ④ Inferences concerning $\mu_{Y|x}$ and the prediction of future Y values

Simple Linear Regression

What is it? : Regression analysis deals with the relationships between two or more , usually continuous, variables.

Model : The simple linear regression model assumes a relationship between observations of a response, $y_i, i = 1, \dots, n$, and the corresponding predictor observations, $x_i, i = 1, \dots, n$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n$$

Parameters : The parameters to be estimated are β_0 , the intercept, and β_1 , the slope.

We also estimate σ^2 , the variance of the noise term.

Simple Linear Regression

- Observe that given $X = x$,

$$E(Y) = E(\beta_0 + \beta_1 x + \varepsilon)$$

$$= \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

$$V(Y) = V(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

- Observed data:

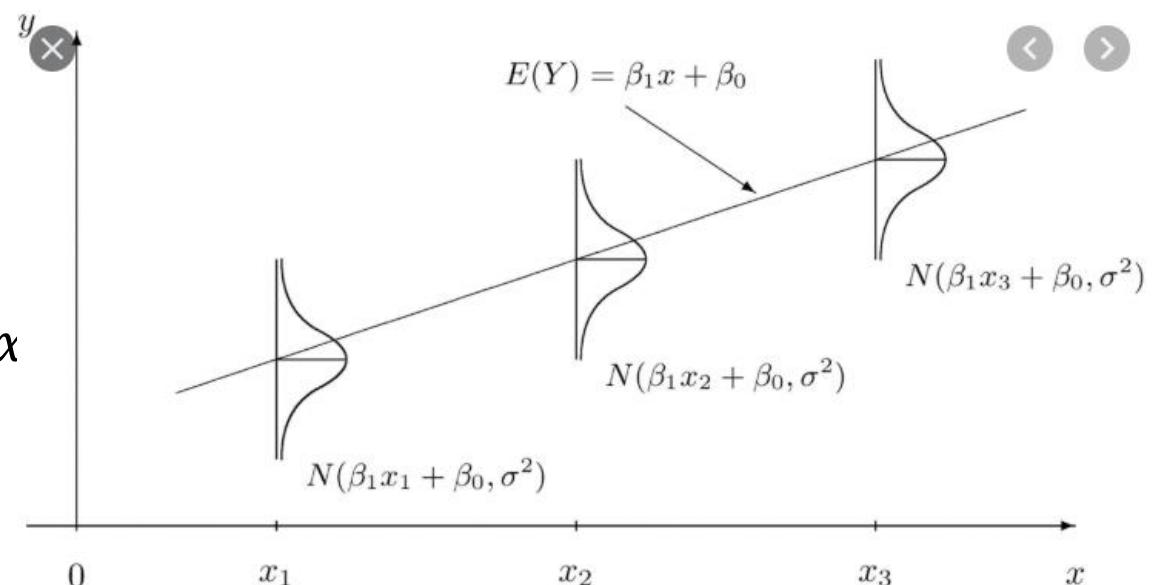
$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

- n observations:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed.

- First question: How to find good estimators of β_0 and β_1 ?



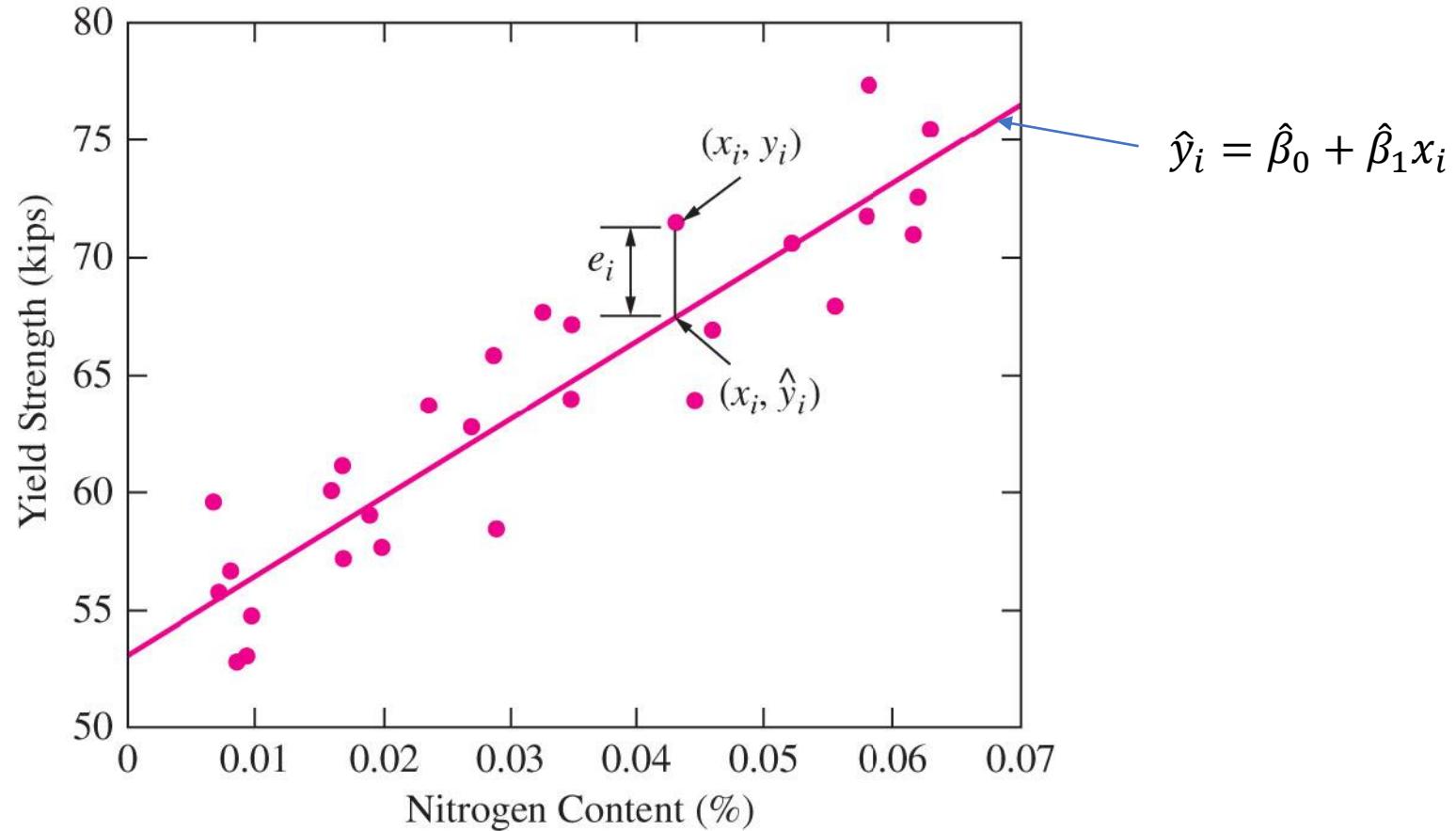


Figure 2.7 Yield strength versus nitrogen content for a sample of welds

Method of Least Squares (MLS)

- <https://mathworld.wolfram.com/LeastSquaresFitting.html>

$$\sum e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Consider a function of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$R^2(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$$\frac{\partial(R^2)}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$$

$$\sum_{i=1}^n y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i = 0$$

$$\frac{\partial(R^2)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] x_i = 0$$

$$\sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

- This leads to equation

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$\hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Derivation of $\hat{\beta}_0$ and $\hat{\beta}_1$

- In matrix form

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

- So

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \quad \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \right)$$

- The 2×2 matrix inverse is

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$= \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{bmatrix}$$

Derivation of $\hat{\beta}_0$ and $\hat{\beta}_1$

- So

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (\bar{x} = \frac{\sum_{i=1}^n x_i}{n})$$

$$= \frac{\bar{y} \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$= \frac{(\sum_{i=1}^n x_i y_i) - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\begin{aligned} (\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})) &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} = (\sum_{i=1}^n x_i y_i) - n \bar{x} \bar{y} \end{aligned}$$

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Method of Least Squares (MLS)

- Let

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Solve the MLS equations, we obtain that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$:
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimator for β_0 and β_1 respectively.

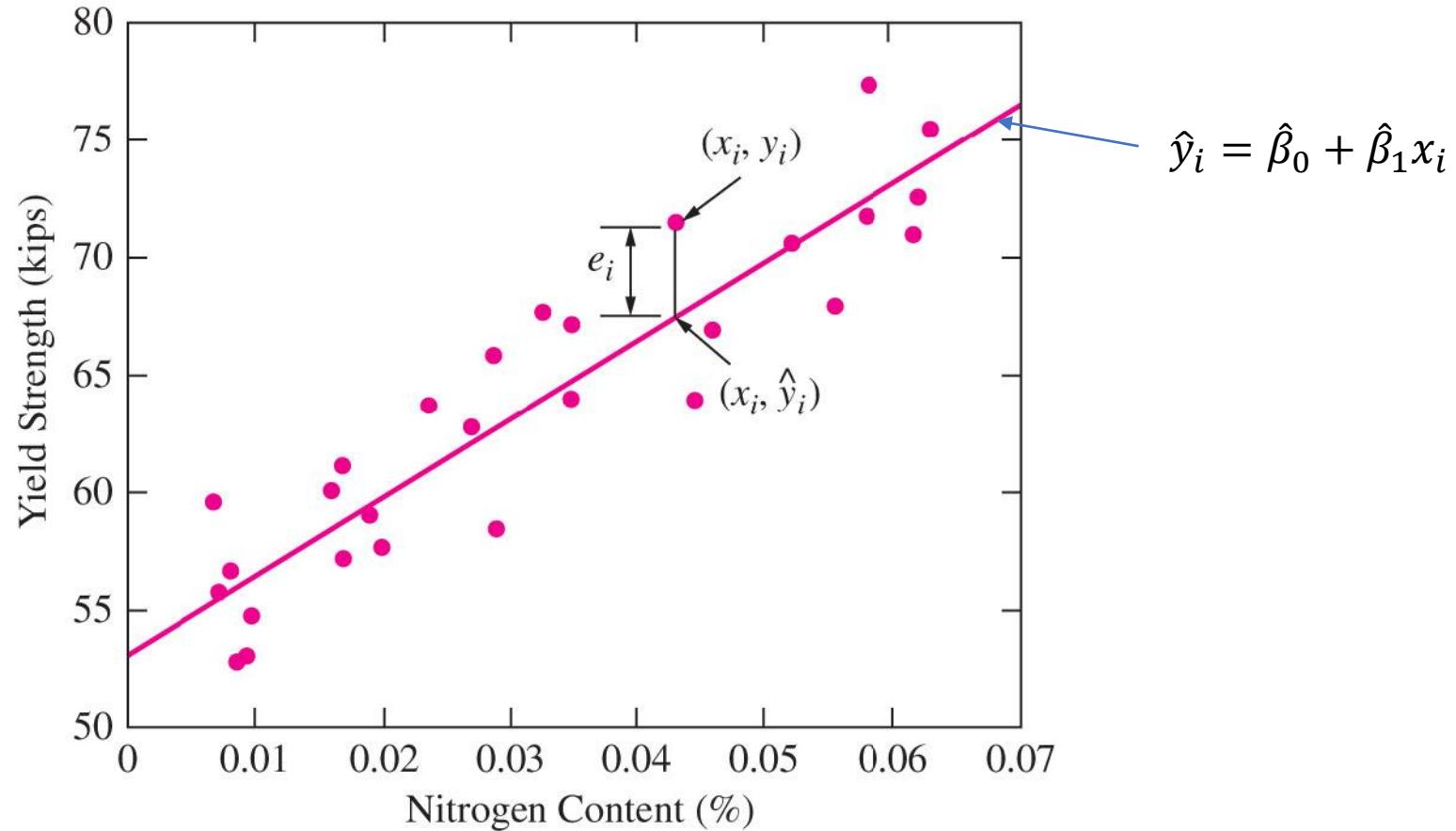


Figure 2.7 Yield strength versus nitrogen content for a sample of welds

Method of Least Squares (MLS)

- <https://mathworld.wolfram.com/LeastSquaresFitting.html>

$$\min \sum e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Let

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Solve the MLS equations, we obtain that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Example 12.4

The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine.

Determination of this number for a biodiesel fuel is expensive and time-consuming.

Therefore a way of predicting this number is wanted.

The data on the table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels.

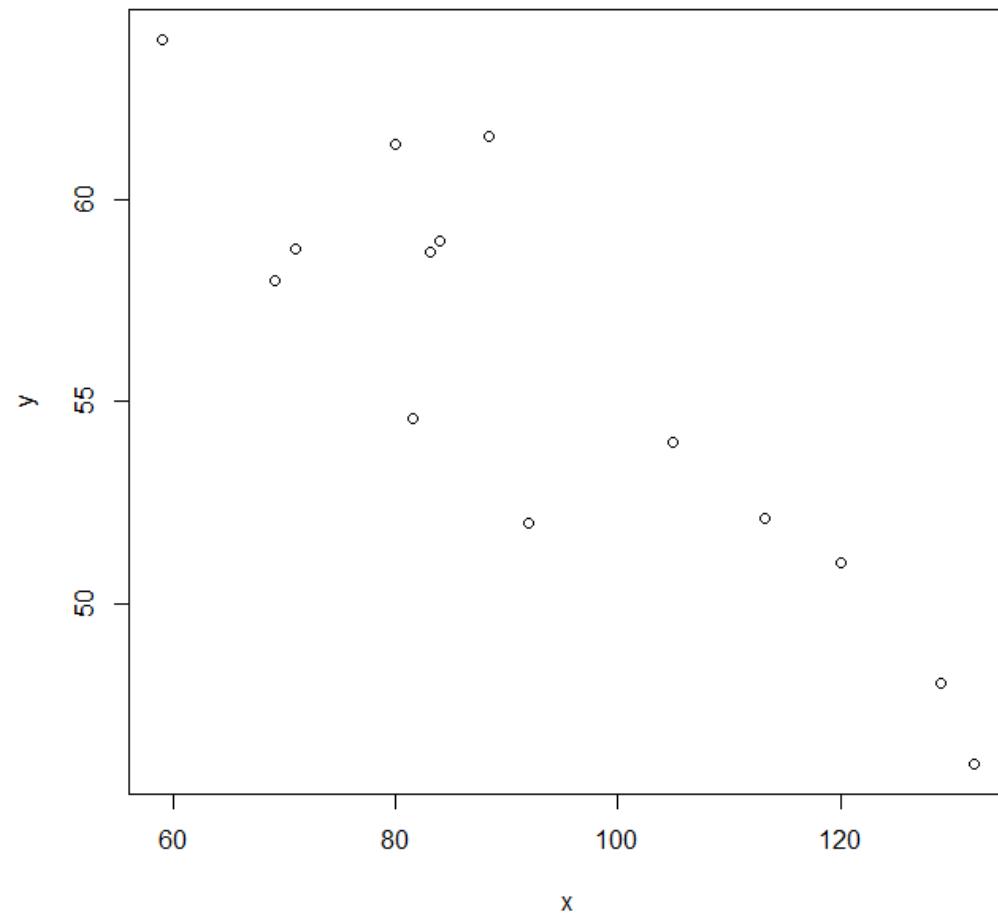
The iodine value is the amount of iodine necessary to saturate a sample of 100g of oil.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	
	61.4	54.6	58.8	58.0							

```
> x <- c(132.0, 129.0, 120.0, 113.2, 105.0, 92.0, 84.0, 83.2, 88.4, 59.0, 80.0, 81.5, 71.0, 69.2)
```

```
> y <- c(46.0, 48.0, 51.0, 52.1, 54.0, 52.0, 59.0, 58.7, 61.6, 64.0, 61.4, 54.6, 58.8, 58.0)
```

```
> f <- lm(y~x)
```



Steps in the analysis

- a) What is the estimated regression line?

$$n = 14$$

$$\bar{x} = 93.393$$

$$\bar{y} = 55.657$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 6802.7693$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = -1424.41429$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -0.2094$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 55.657 - (-0.2094)(93.393) = 75.212$$

The equation of the estimated regression line is $y = 75.212 - 0.2094 x$.

Steps in the analysis

```
> f <- lm(y ~ x)  
> summary(f)
```

Call:

lm(formula = y ~ x)

Residuals:

	Min	1Q	Median	3Q	Max
	-3.9488	-1.5665	0.6817	1.0846	4.8974

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	75.21243	2.98363	25.208	9.22e-12 ***
x	-0.20939	0.03109	-6.734	2.09e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.564 on 12 degrees of freedom

Multiple R-squared: 0.7908, Adjusted R-squared: 0.7733

F-statistic: 45.35 on 1 and 12 DF, p-value: 2.091e-05

Residual Sum of Squares

- How to estimate σ^2 ?
- Residuals : $e_i = y_i - \hat{y}_i, i = 1, \dots, n$
- The residual sum of squares (error sum of squares) :

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

- Estimate of σ^2 : $\hat{\sigma}^2 = SSE/(n - 2)$

Example 12.4

The data on the next table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels. Find the estimate of σ .

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4
	54.6	58.8	58.0								
Fit	47.57	48.20	50.09	51.51	53.23	55.95	57.62	57.79	56.70	62.86	58.46
	58.15	60.35	60.73								
Residual	-1.57	-0.20	0.91	0.59	0.77	-3.95	1.34	0.91	4.90	1.14	2.94
	-3.55	-1.55	-2.725								

$$\hat{y} = 75.212 - 0.2094 x$$

```
> res <- y -(75.21243-0.20939*x)  
  
> sum(res^2)/(length(x)-2) # 6.576655  
  
> sqrt(sum(res^2)/(length(x)-2)) # 2.564499
```

Simple Linear Regression

Model : The simple linear regression model assumes a relationship between observations of a response, $y_i, i = 1, \dots, n$, and the corresponding predictor observations, $x_i, i = 1, \dots, n$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n$$

Parameters : The parameters to be estimated are β_0 , the intercept, and β_1 , the slope.

We also estimate σ^2 , the variance of the noise term.

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = SSE/(n - 2) \quad (s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1})$$

$$\text{where } SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \quad E(Y_i) = \beta_0 + \beta_1 x_i$$

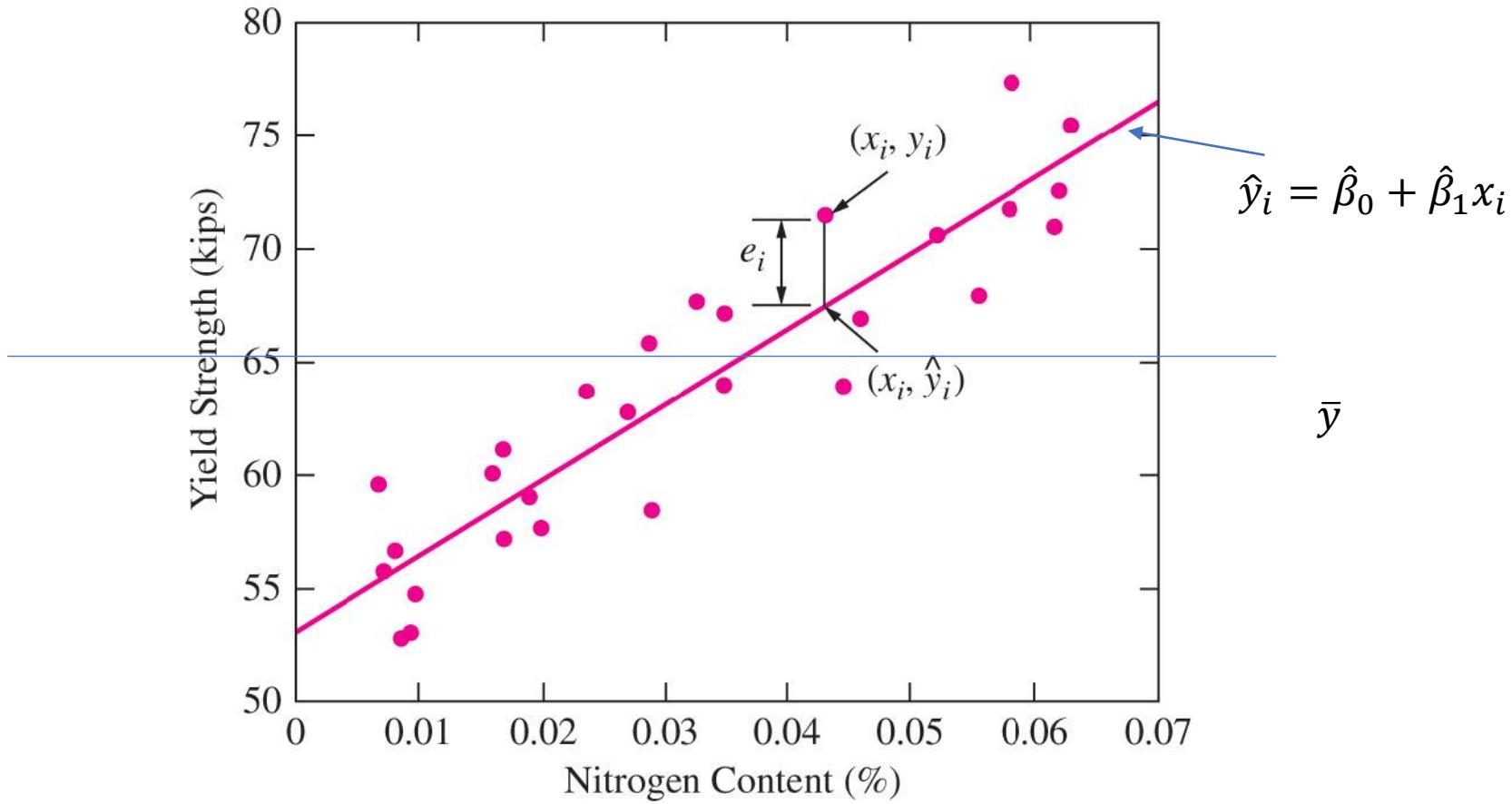


Figure 2.7 Yield strength versus nitrogen content for a sample of welds

Coefficient of determination

The coefficient of determination, R^2 or r^2 , is the proportion of the variation in the response "explained" by the predictor.

$$r^2 = 1 - \frac{SSE}{SST}, 0 \leq r^2 \leq 1$$

$$\text{where } SST = S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - (\sum y_i)^2/n$$

A high r^2 (i.e. close to 1) indicates a successful model in the sense that the residual variability is much smaller than the original variability.

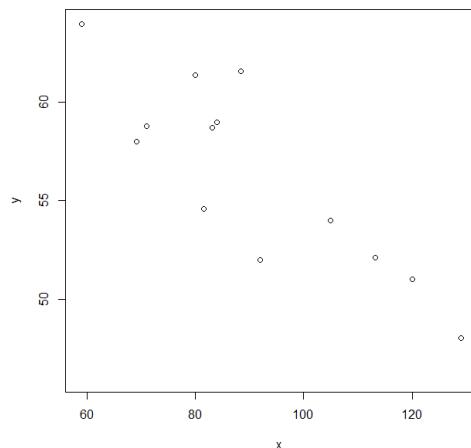
Example 12.9. The scatter plot of the iodine value-cetane number data shows a reasonably high r^2 value.

```
SST :> sum((y-mean(y))^2) # 377.1743
```

```
SSE :> sum(res^2) # 78.91986
```

The coefficient of determination is then

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{78.920}{377.1743} = 0.791$$



Steps in the analysis

```
> f <- lm(y ~ x)  
> summary(f)
```

Call:

lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-3.9488	-1.5665	0.6817	1.0846	4.8974

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	75.21243	2.98363	25.208	9.22e-12 ***
x	-0.20939	0.03109	-6.734	2.09e-05 ***

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	'.'	0.1 ''	1	

Residual standard error: 2.564 on 12 degrees of freedom

Multiple R-squared: 0.7908, Adjusted R-squared: 0.7733

F-statistic: 45.35 on 1 and 12 DF, p-value: 2.091e-05

12.3 Inferences About the Slope Parameter β_1

The values of x_i 's are assumed to be chosen before the experiment is performed, so only the Y_i 's are random.

The estimators for β_0 and β_1 are obtained by replacing y_i by Y_i .

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n}$$

The denominator of $\hat{\beta}_1$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, depends only on the x_i 's and not on the Y_i 's, so it is a constant.

Because $\sum_{i=1}^n (x_i - \bar{x}) \bar{Y} = \bar{Y} \sum_{i=1}^n (x_i - \bar{x}) = \bar{Y} \cdot 0 = 0$, the slope estimator can be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}} = \sum_{i=1}^n c_i Y_i, \text{ where } c_i = (x_i - \bar{x}) / S_{xx}$$

12.3 Inferences About the Slope Parameter β_1

Proposition

1. The mean value of $\hat{\beta}_1$: $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 .
2. The variance and standard deviation of $\hat{\beta}_1$:

$$V(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}}, \quad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

Replacing σ by its estimate s gives an estimate of $\sigma_{\hat{\beta}_1}$.

3. The estimator $\hat{\beta}_1$ has a normal distribution (because it is a linear function of independent normal rv's).

cf)<https://stats.stackexchange.com/questions/12186/expected-value-and-variance-of-estimation-of-slope-parameter-beta-1-in-simple>

Theorem

The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

has a t distribution with $(n - 2)$ df.

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}\right) = E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{S_{xx}}\right) \\
&= E\left(\frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right) \\
&= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i}{S_{xx}}\right) = \beta_1 E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{S_{xx}}\right) = \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{S_{xx}} = \beta_1
\end{aligned}$$

Here $\sum_{i=1}^n (x_i - \bar{x}) x_i = \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) \bar{x} = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = S_{xx}$

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}\right) = Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{S_{xx}}\right) \\
&= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_0}{S_{xx}} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i}{S_{xx}} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right) \\
&= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right) = Var\left(\frac{(x_1 - \bar{x}) \varepsilon_1}{S_{xx}} + \dots + \frac{(x_n - \bar{x}) \varepsilon_n}{S_{xx}}\right) \\
&= \frac{(x_1 - \bar{x})^2 \sigma^2}{(S_{xx})^2} + \dots + \frac{(x_n - \bar{x})^2 \sigma^2}{(S_{xx})^2} = \sigma^2 \left[\frac{\sum (x_i - \bar{x})^2}{(S_{xx})^2} \right] = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

A Confidence Interval for β_1

We begin with a probability statement

$$P\left(-t_{\frac{\alpha}{2}, n-2} < \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} < t_{\frac{\alpha}{2}, n-2}\right) = 1 - \alpha$$

A $100(1-\alpha)\%$ CI for the slope β_1 of the true regression line is

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{\beta}_1}$$

$$-t_{\frac{\alpha}{2}, n-2} < \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} < t_{\frac{\alpha}{2}, n-2}$$

$$-t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} < \hat{\beta}_1 - \beta_1 < t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$$

$$-\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} < -\beta_1 < -\hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$$

$$\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$$

Example

The data on the next table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels. Compute the 95% CI for the slope β_1 .

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4
	54.6	58.8	58.0								

$$\begin{aligned}\hat{\beta}_1 \pm t_{0.025,12} \cdot s_{\hat{\beta}_1} &= \hat{\beta}_1 \pm t_{0.025,12} \cdot \frac{s}{\sqrt{s_{xx}}} \\ &= -0.20939 \pm 2.179 \times \frac{2.564}{82.4789} \\ &= -0.20939 \pm 2.179 \times 0.03109 \rightarrow (-0.2771, 0.1416)\end{aligned}$$

Steps in the analysis

```
> f <- lm(y ~ x)
> summary(f)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.9488	-1.5665	0.6817	1.0846	4.8974

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	75.21243	2.98363	25.208	9.22e-12 ***
x	-0.20939	0.03109	-6.734	2.09e-05 ***

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	'.'	0.1 ''	1	

Residual standard error: 2.564 on 12 degrees of freedom

Multiple R-squared: 0.7908, Adjusted R-squared: 0.7733

F-statistic: 45.35 on 1 and 12 DF, p-value: 2.091e-05

Hypothesis-Testing Procedures

Null hypothesis : $H_0: \beta_1 = \beta_{10}$

Test statistic value : $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Alternative Hypothesis

$$H_a: \beta_1 > \beta_{10}$$

$$H_a: \beta_1 < \beta_{10}$$

$$H_a: \beta_1 \neq \beta_{10}$$

Rejection Region for Level α Test

$$t \geq t_{\alpha, n-2}$$

$$t \leq -t_{\alpha, n-2}$$

$$\text{either } t \geq t_{\alpha/2, n-2} \text{ or } t \leq -t_{\alpha/2, n-2}$$

A P -value based on $n - 2$ df can be calculated just as was done previously.

The model utility test is the test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, in which case the test statistic value is the t ratio of $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.

Example

The data on the next table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels. Find the P -value for $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4
	54.6	58.8	58.0								

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{-0.20939}{0.03109} = -6.734$$

$$P\text{-value} = P(t_{14-2} < -6.734) + P(t_{14-2} > 6.734) = 2 \times P(t_{14-2} < -6.734) = 2.091352e-05$$

This value is the same as the value at the column of $\Pr(>|t|)$ on page 10.

```
> 2*pt(-6.734, 12)
```

```
[1] 2.091352e-05
```

Since $t = -6.734 \leq -t_{0.025,12} = -2.179$, H_0 can be rejected under the level 0.05 test.

Inferences Concerning $\mu_{Y|x^*}$ and the Prediction of Future Y Values

Proposition

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is some fixed value of x . Then

1. The mean value of \hat{Y} is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

2. The variance of \hat{Y} is

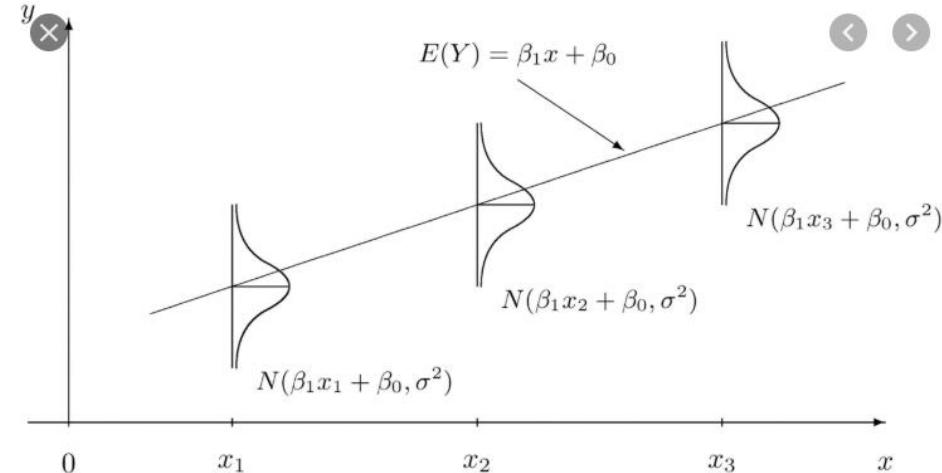
$$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

3. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ results from replacing σ by its estimate s .

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

A $100(1-\alpha)\%$ CI for $\mu_{Y|x^*}$, the expected value of Y when $x = x^*$, is $\frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{y}}$$



$$V(\hat{Y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) = Var(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x^*) = Var(\bar{y} + \hat{\beta}_1(x^* - \bar{x}))$$

$$= Var(\bar{y}) + (x^* - \bar{x})^2 Var(\hat{\beta}_1) + 2(x^* - \bar{x}) Cov(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2(x^* - \bar{x})^2}{S_{xx}} + 0$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{\sum y_i}{n}, \sum c_i y_i\right) \quad (c_i = (x_i - \bar{x})/S_{xx})$$

$$= \frac{1}{n} \sum c_i Var(y_i)$$

$$= \frac{\sigma^2}{n} \sum c_i = 0$$

> <https://math.stackexchange.com/questions/2912624/show-that-cov-bary-hat-beta-1-0>

A Prediction Interval of Future Y Values

Consider some future observation value Y when the independent variable has value x^* .

The error of prediction is $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, a difference between two random variables.

Because the future value Y is independent of the observed Y_i 's,

$$\begin{aligned} V(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)) &= V(Y) + V(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] \end{aligned}$$

A $100(1-\alpha)\%$ CI for a future observation Y to be made when $x = x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Example

The data on the table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels.

- Compute the 95% CI for $\mu_{Y|100}$, the expected value of Y when $x = 100$

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	
	61.4	54.6	58.8	58.0							

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 100 &\pm t_{0.025,12} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{\beta}_0 + \hat{\beta}_1 100 \pm t_{0.025,12} \cdot s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \\ &= 75.21243 + (-0.20939) \times 100 \pm 2.179 \cdot 2.564 \sqrt{\frac{1}{14} + \frac{(100 - 93.393)^2}{6802.769}} \rightarrow (52.715, 55.832) \end{aligned}$$

> predict(f, list(x=c(100)), interval="conf")

- Compute the 95% CI for a future observation Y when $x = 100$ is

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 100 &\pm t_{0.025,12} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{\beta}_0 + \hat{\beta}_1 100 \pm t_{0.025,12} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \\ &= 75.21243 + (-0.20939) \times 100 \pm 2.179 \cdot 2.564 \sqrt{1 + \frac{1}{14} + \frac{(100 - 93.393)^2}{6802.769}} \rightarrow (48.473, 60.074) \end{aligned}$$

> predict(f, list(x=c(100)), interval="pred")

12.5 Correlation

The Sample Correlation Coefficient for the n pairs $(x_1, y_1), \dots, (x_n, y_n)$ is

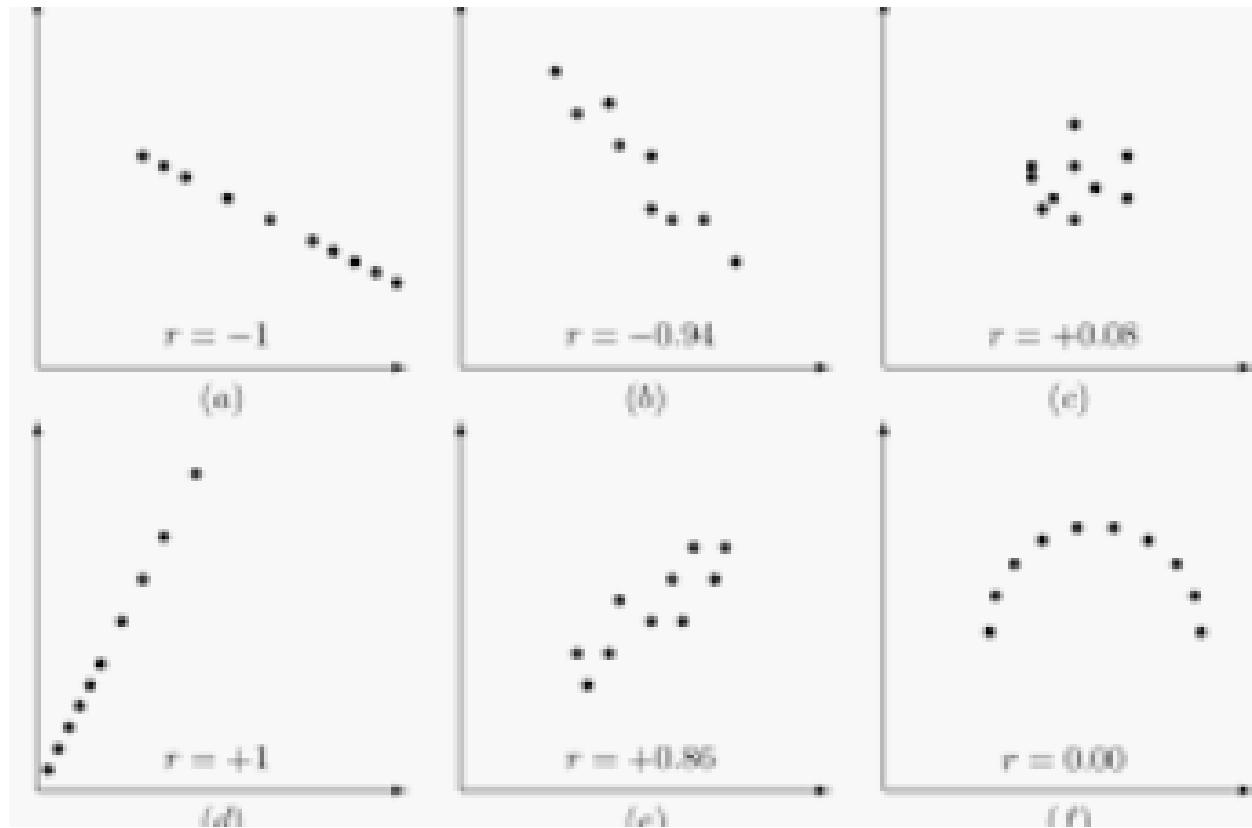
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx}} \sqrt{s_{yy}}}$$

Properties of r

1. The value of r is independent of the units in which x and y are measured.
2. $-1 \leq r \leq 1$
3. $r = 1$ if and only if all (x_i, y_i) pairs lie on a straight line with positive slope and

$r = -1$ if and only if all (x_i, y_i) pairs lie on a straight line with negative slope.

5. The square of the sample correlation coefficient gives the value of the coefficient of determination.



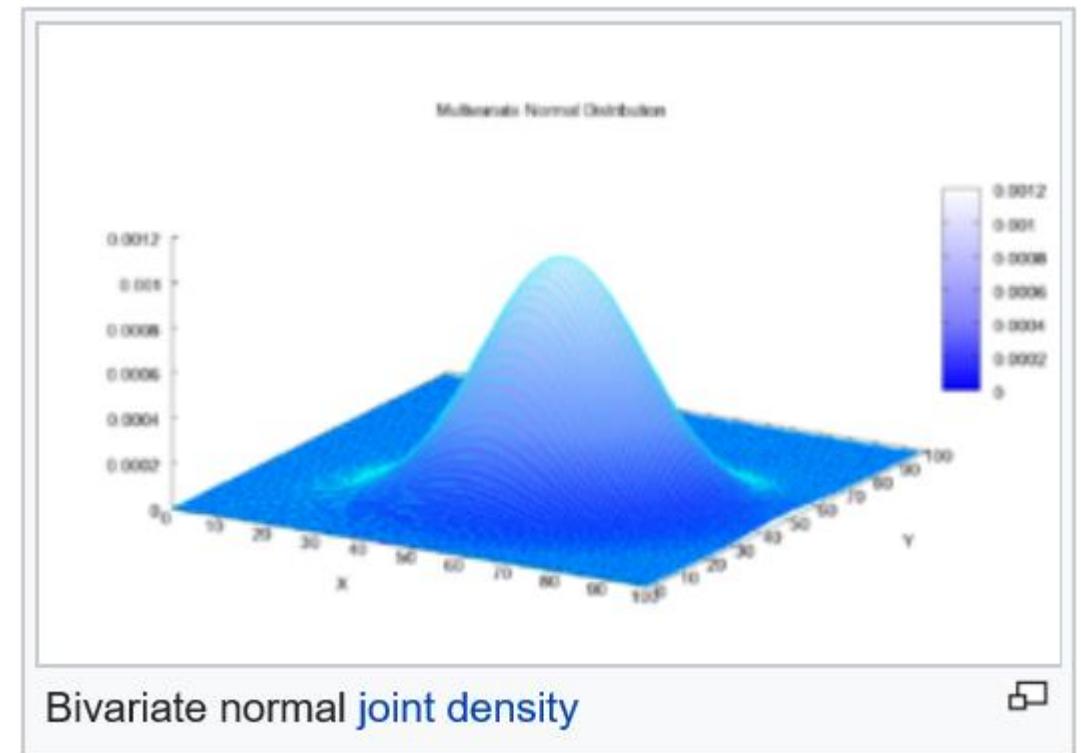
12.5 Correlation

Assumption

The joint probability distribution of (X, Y) is specified by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]/[2(1-\rho^2)]}$$
$$-\infty < X < \infty, -\infty < Y < \infty$$

$f(x, y)$ is called the bivariate normal probability distribution.



Testing for the Absence of Correlation

When : $H_0 : \rho = 0$ is true, the test statistic

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

has a t distribution with $n - 2$ df.

Alternative Hypothesis

$$H_a: \rho > 0$$

$$H_a: \rho < 0$$

$$H_a: \rho \neq 0$$

Rejection Region for Level α Test

$$t \geq t_{\alpha, n-2}$$

$$t \leq -t_{\alpha, n-2}$$

$$\text{either } t \geq t_{\alpha/2, n-2} \text{ or } t \leq -t_{\alpha/2, n-2}$$

A P -value based on $n - 2$ df can be calculated as described previously.

- The relationship between the diameter of a nail(x) and its withdrawal strength(y)

x	2.52	2.87	3.05	3.43	3.68	3.76	3.76	4.50	4.50	5.26
y	54.74	59.01	72.92	50.85	54.99	60.56	69.08	77.03	69.97	90.70

- $H_0 : \rho = 0$ vs $H_a : \rho > 0$

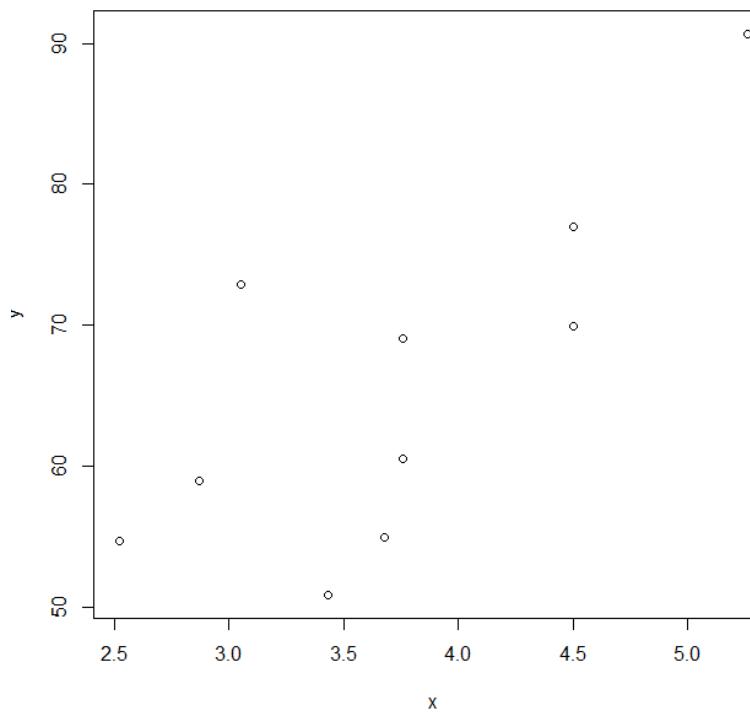
$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7492\sqrt{10-2}}{\sqrt{1-0.7492^2}} = 3.199$$

$$\text{p-value} = P(t_8 > 3.199) = 0.00632$$

Since p-value = 0.00632 < 0.01, it is reasonable to conclude that $\rho > 0$.

```
> 1-pt(3.199, 8) # 0.00632
```

```
> x <- c(2.52, 2.87, 3.05, 3.43, 3.68, 3.76, 3.76, 4.50, 4.50, 5.26)
> y <- c(54.74, 59.01, 72.92, 50.85, 54.99, 60.56, 69.08, 77.03, 69.97, 90.70)
> plot(x, y)
> cor(x, y)
[1] 0.7491725
```



Chapter 10 – The Analysis of Variance

Outline

- ① Single-Factor ANOVA
- ② Multiple Comparisons in ANOVA
- ③ More on Single-Factor ANOVA

10.1 Single Factor ANOVA

- t-test : used to compare means of two populations
- ANONA(Analysis of Variance) : used to compare more than two means.

I : the number of populations or treatments being compared

μ_i : the mean of population i

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

$$H_a: \text{at least two of the } \mu_i \text{'s are different}$$

The independent variables are termed the factor or treatment, and the various categories within that treatment are termed the levels.

10.1 Single Factor ANOVA Example 10.1

We want to compare 4 types of boxes with respect to the compression strength.

We obtain 6 observations for each type of box.

Type of Box	Compression Strength (lb)							Sample Mean	Sample SD
1	655.5	788.3	734.3	721.4	679.1	699.4		713.00	46.55
2	789.2	772.5	786.9	686.1	732.1	774.8		756.93	40.34
3	737.1	639.0	696.3	671.7	717.2	727.1		698.07	37.20
4	535.1	628.7	542.4	559.0	586.9	520.0		562.02	39.87
					Grand mean =		682.50		

We want to test if all four types have the same mean compression strength.

How to construct the appropriate hypotheses? How to perform the test?

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a: \text{at least two of the } \mu_i \text{'s are different}$$

11.1 Two Factor ANOVA Example 11.1

Compare three different brands of erasable pens and four different wash treatments with respect to their ability to remove marks on a fabric.

The data shows the overall specimen color change; the lower this value, the more marks were removed.

Brand of Pen	Washing Treatment				Total	Average
	1	2	3	4		
1	0.97	0.48	0.48	0.46	2.39	0.598
	0.77	0.14	0.22	0.25	1.38	0.345
	0.67	0.39	0.57	0.19	1.82	0.455
Total	2.41	1.01	1.27	0.90	Grand average : 0.466	
Average	0.803	0.337	0.423	0.300		

10.1 Single Factor ANOVA Example 10.1

Suppose that we want to test if first and second type of box have the same mean compression strength.

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_a: \mu_1 - \mu_2 \neq 0$$

```
> type1 <- c(655.5,788.3,734.3,721.4,679.1,699.4)
> type2 <- c(789.2,772.5,786.9,686.1,732.1,774.8)
> t.test(type1, type2)

  Welch Two Sample t-test

data: type1 and type2
t = -1.7471, df = 9.8014, p-value = 0.1118
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-100.11614 12.24947
sample estimates:
mean of x mean of y
713.0000 756.9333
```

Since $p\text{-value}=0.1118 > 0.05$, we do not reject H_0 . We do not have enough evidence to say that the means of the first and second type of box are different.

R execution of Example 10.1

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs $H_a:$ at least two of the μ_i 's are different

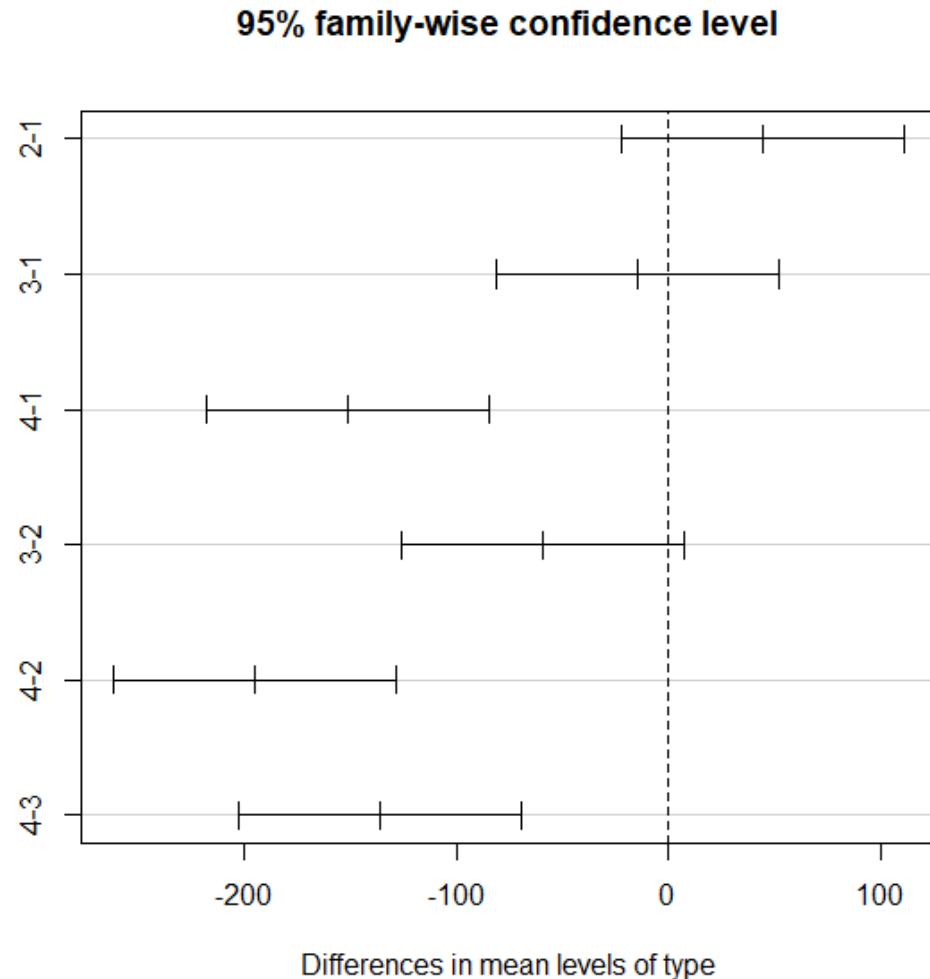
```
> compression <- c(655.5,788.3,734.3,721.4,679.1,699.4,  
+ 789.2,772.5,786.9,686.1,732.1,774.8,  
+ 737.1,639.0,696.3,671.7,717.2,727.1,  
+ 535.1,628.7,542.4,559.0,586.9,520.0)  
> type <- c(rep("1", 6), rep("2", 6), rep("3", 6), rep("4", 6))  
> oneway <- aov(compression ~ type)  
> summary(oneway)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	127375	42458	25.09	5.53e-07 ***
Residuals	20	33839	1692		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Since p-value= 5.53e-07 < 0.05, we can reject H_0 . We can say that at least two of the μ_i 's are different

```
> posthoc <- TukeyHSD(x=oneway, 'type',  
conf.level=0.95)  
> posthoc  
> plot(posthoc)
```



ANOVA: notation (equal sample size)

X_{ij} : j^{th} measurement taken from the i^{th} population

sample sizes : J for each population

Individual sample mean : $\bar{X}_{1.}, \bar{X}_{2.}, \dots, \bar{X}_{I.}$

$$\bar{X}_{i.} = \frac{\sum_{j=1}^J X_{ij}}{J}, i = 1, 2, \dots, I$$

Grand mean : $\bar{X}_{..}$

$$\bar{X}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^J X_{ij}}{IJ}$$

Individual sample variance : $S_1^2, S_2^2, \dots, S_I^2$

$$S_i^2 = \frac{\sum_{j=1}^J (X_{ij} - \bar{X}_{i.})^2}{J-1}$$

10.1 Single Factor ANOVA Example 10.1

We want to compare 4 types of boxes with respect to the compression strength.

We obtain 6 observations for each type of box.

Type of Box	Compression Strength (lb)							Sample Mean	Sample SD
1	655.5	788.3	734.3	721.4	679.1	699.4		713.00	46.55
2	789.2	772.5	786.9	686.1	732.1	774.8		756.93	40.34
3	737.1	639.0	696.3	671.7	717.2	727.1		698.07	37.20
4	535.1	628.7	542.4	559.0	586.9	520.0		562.02	39.87
					Grand mean =		682.50		

We want to test if all four types have the same mean compression strength.

How to construct the appropriate hypotheses? How to perform the test?

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a: \text{at least two of the } \mu_i \text{'s are different}$$

ANOVA

- Assumption
 1. All samples are independent of each other.
 2. Each population or treatment distributions are normal with $E(X_{ij}) = \mu_i$.
 3. Each population has the same variance $\text{Var}(X_{ij}) = \sigma^2$.

ANOVA

- The Test Statistic :

Mean square for treatments is given by

$$\begin{aligned}MSTr &= \frac{\sum_{i=1}^I \sum_{j=1}^J (\bar{X}_{i\cdot} - \bar{X}_{..})^2}{I-1} \\&= \frac{J}{I-1} [(\bar{X}_{1\cdot} - \bar{X}_{..})^2 + (\bar{X}_{2\cdot} - \bar{X}_{..})^2 + \dots + (\bar{X}_{I\cdot} - \bar{X}_{..})^2] \\&= \frac{J}{I-1} \sum_i (\bar{X}_{i\cdot} - \bar{X}_{..})^2\end{aligned}$$

Mean square for error is

$$MSE = \frac{s_1^2 + s_2^2 + \dots + s_I^2}{I}$$

The Test Statistic for single-factor ANOVA is

$$F = MSTr / MSE$$

ANOVA

- Notice that
 1. The quantities $MStr$ and MSE are statistic (random quantities that can be computed based on samples)
 2. Each S_i^2 assesses variation within a particular sample. So MSE is a measure of *within-samples* variation.
 3. The quantity $MStr$ measures how much similar or different the samples are in terms of their mean. This is a measure of *between sample* variation.
 4. Thus the test statistic F compares within sample variation to the between sample variation.

ANOVA

Proposition

When $H_0: \mu_1 = \mu_2 = \dots = \mu_I$ is true,

$$E(MStr) = E(MSE) = \sigma^2$$

whereas when $H_0: \mu_1 = \mu_2 = \dots = \mu_I$ is false,

$$E(MStr) > E(MSE) = \sigma^2$$

That is, both statistics are unbiased for estimating the common population variance σ^2 when H_0 is true, but $MStr$ tends to overestimate σ^2 when H_0 is false.

Inference Concerning Two Population Variances

- $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_a : \sigma_1^2 > \sigma_2^2$
- The F Distribution

The F probability distribution has two parameters, denoted by ν_1 and ν_2 .

The parameter ν_1 is called the *numerator degrees of freedom*, and ν_2 is the *denominator degrees of freedom*.

If X_1 and X_2 are independent chi-squared rv's with ν_1 and ν_2 df, respectively, then the rv

$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

(the ratio of the two chi-squared variables divided by their respective degrees of freedom), can be shown to have an F distribution.

ANOVA

ANONA(Analysis of Variance) : used to compare more than two means.

Proposition

When $H_0: \mu_1 = \mu_2 = \dots = \mu_I$ is true,

$$E(MStr) = E(MSE) = \sigma^2 \quad X_{ij} \sim N(\mu_i, \sigma^2)$$

whereas when $H_0: \mu_1 = \mu_2 = \dots = \mu_I$ is false,

$$E(MStr) > E(MSE) = \sigma^2$$

That is, both statistics are unbiased for estimating the common population variance σ^2 when H_0 is true, but $MStr$ tends to overestimate σ^2 when H_0 is false.

$$MStr = \frac{\sum_{i=1}^I \sum_{j=1}^J (\bar{X}_{i\cdot} - \bar{X}_{..})^2}{I-1} = \frac{J \sum_{i=1}^I (\bar{X}_{i\cdot} - \bar{X}_{..})^2}{I-1} = \frac{SSTr}{I-1} \quad (\sum_{j=1}^J a = Ja)$$

$$MSE = \frac{s_1^2 + s_2^2 + \dots + s_I^2}{I}$$

$$\bar{X}_{i\cdot} = \frac{\sum_{j=1}^J X_{ij}}{J}, \quad \bar{X}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^J X_{ij}}{IJ}$$

10.1 Single Factor ANOVA Example 10.1

We want to compare 4 types of boxes with respect to the compression strength.

We obtain 6 observations for each type of box.

Type of Box	Compression Strength (lb)							Sample Mean	Sample SD
1	655.5	788.3	734.3	721.4	679.1	699.4		713.00	46.55
2	789.2	772.5	786.9	686.1	732.1	774.8		756.93	40.34
3	737.1	639.0	696.3	671.7	717.2	727.1		698.07	37.20
4	535.1	628.7	542.4	559.0	586.9	520.0		562.02	39.87
					Grand mean =		682.50		

We want to test if all four types have the same mean compression strength.

How to construct the appropriate hypotheses? How to perform the test?

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a: \text{at least two of the } \mu_i \text{'s are different}$$

Example 10.2 (Example 10.1 continued) :

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs $H_1 : \text{At least two } \mu_i\text{'s are different}$

$$\bar{x}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^J x_{ij}}{IJ} = 682.50$$

$$MStr = \frac{6}{4-1} [(713.00 - 682.50)^2 + (756.93 - 682.50)^2 + (698.07 - 682.50)^2 + (562.02 - 682.50)^2] = 42455.86$$

$$MSE = \frac{1}{4} [(46.55)^2 + (40.34)^2 + (37.20)^2 + (39.87)^2] = 1691.92$$

$$f = MStr / MSE = \frac{42455.86}{1691.92} = 25.09$$

Since $25.09 \geq 3.10 = F_{0.05,3,20}$, H_0 is rejected at significance level 0.05.

$$\nu_1 = I - 1 = 4 - 1 = 3, \nu_2 = I(J - 1) = 4(6 - 1) = 20$$

> qf(0.95, 3, 20)

[1] 3.098391

Sum of Squares

$$\begin{aligned}SSTr &= \sum_{i=1}^I \sum_{j=1}^J (\bar{X}_{i\cdot} - \bar{X}_{..})^2 = J \sum_{i=1}^I (\bar{X}_{i\cdot} - \bar{X}_{..})^2 \quad (\sum_{j=1}^J a = Ja) \\&= J \sum_{i=1}^I \left(\bar{X}_{i\cdot}^2 - 2\bar{X}_{..}\bar{X}_{i\cdot} + \bar{X}_{..}^2 \right) = J \left\{ \sum_{i=1}^I \left(\frac{\bar{X}_{i\cdot}}{J} \right)^2 - 2 \frac{\bar{X}_{..}}{IJ} \sum_{i=1}^I \frac{\bar{X}_{i\cdot}}{J} + I \left(\frac{\bar{X}_{..}}{IJ} \right)^2 \right\} \\&= J \left\{ \sum_{i=1}^I \left(\frac{\bar{X}_{i\cdot}}{J} \right)^2 - 2 \frac{\bar{X}_{..}}{IJ} \frac{\bar{X}_{..}}{J} + I \left(\frac{\bar{X}_{..}}{IJ} \right)^2 \right\} \\&= J \left\{ \sum_{i=1}^I \left(\frac{\bar{X}_{i\cdot}}{J} \right)^2 - 2I \frac{\bar{X}_{..}}{IJ} \frac{\bar{X}_{..}}{IJ} + I \left(\frac{\bar{X}_{..}}{IJ} \right)^2 \right\} \\&= J \left\{ \sum_{i=1}^I \left(\frac{\bar{X}_{i\cdot}}{J} \right)^2 - I \left(\frac{\bar{X}_{..}}{IJ} \right)^2 \right\} = \frac{1}{J} \sum_i X_{i\cdot}^2 - \frac{1}{IJ} X_{..}^2\end{aligned}$$

$E(MStr) = \sigma^2$ if H_0 is true

For any random variable $E[Y^2] = V(Y) + (E[Y])^2$

$$\begin{aligned} E(SStr) &= E\left[\frac{1}{J}\sum_i X_{i.}^2 - \frac{1}{IJ}X_{..}^2\right] = \frac{1}{J}\sum_i E(X_{i.}^2) - \frac{1}{IJ}E(X_{..}^2) \quad [E(aX + bY) = aE(X) + bE(Y)] \\ &= \frac{1}{J}\sum_i\{V(X_{i.}) + [E(X_{i.})]^2\} - \frac{1}{IJ}\{V(X_{..}) + [E(X_{..})]^2\} \quad X_{i.} = \sum_{j=1}^J X_{ij} \quad X_{..} = \sum_{i=1}^I \sum_{j=1}^J X_{ij} \end{aligned}$$

Since $E(X_{..}) = E(J\sum_{i=1}^I X_{ij}) = J\sum_i E(X_{ij}) = J\sum_i \mu_i = IJ\bar{\mu}$, where $\bar{\mu} = \frac{\sum_i \mu_i}{I}$

$$\begin{aligned} E(SStr) &= \frac{1}{J}\sum_i\{J\sigma^2 + [J\mu_i]^2\} - \frac{1}{IJ}[IJ\sigma^2 + [IJ\bar{\mu}]^2] \\ &= \frac{1}{J}\sum_i\{J\sigma^2 + J^2\mu_i^2\} - \frac{1}{IJ}[IJ\sigma^2 + [IJ\bar{\mu}]^2] \\ &= I\sigma^2 + \sum_i\{J\mu_i^2\} - \sigma^2 - IJ\bar{\mu}^2 \\ &= (I-1)\sigma^2 + \sum_i\{J\mu_i^2\} - IJ\bar{\mu}^2 = (I-1)\sigma^2 + J(\sum_i \mu_i^2 - I\bar{\mu}^2) \\ &= (I-1)\sigma^2 + J\sum_i(\mu_i - \bar{\mu})^2 \\ (\sum_i(\mu_i - \bar{\mu})^2) &= \sum_i \mu_i^2 - 2\bar{\mu} \sum_i \mu_i + I\bar{\mu}^2 = \sum_i \mu_i^2 - 2I\bar{\mu}^2 + I\bar{\mu}^2 = \sum_i \mu_i^2 - I\bar{\mu}^2 \end{aligned}$$

$$E(MStr) = E\left(\frac{SStr}{I-1}\right) = \sigma^2 + \frac{J}{I-1}\sum_i(\mu_i - \bar{\mu})^2$$

If H_0 is true, $\mu_1 = \mu_2 = \dots = \mu_I = \bar{\mu}$ and $\sum_i(\mu_i - \bar{\mu})^2 = 0$. So $E(MStr) = \sigma^2$.

Theoretical Results | Stat 414/415 – Statistics Online – Penn State

<https://online.stat.psu.edu/stat415/lesson/13/13.3>

Number of population : m , the sample size of i^{th} population : n_i

$$SSTr = \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^m n_i \bar{X}_{i.}^2 - n \bar{X}_{..}^2$$

$$E(SSTr) = E \left[\sum_{i=1}^m n_i \bar{X}_{i.}^2 - n \bar{X}_{..}^2 \right] = \left[\sum_{i=1}^m n_i E(\bar{X}_{i.}^2) \right] - n E \left[\bar{X}_{..}^2 \right]$$

Since $E[X^2] = Var(X) + \mu^2$,

$$E(SSTr) = \left[\sum_{i=1}^m n_i \left(\frac{\sigma^2}{n_i} + \mu_i^2 \right) \right] - n \left[\frac{\sigma^2}{n} + \bar{\mu}^2 \right]$$

where $\bar{\mu} = \sum_{i=1}^m n_i \mu_i / n$ and $E(\bar{X}_{..}) = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} E(X_{ij}) = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \mu_i = \frac{1}{n} \sum_{i=1}^m n_i \mu_i = \bar{\mu}$

$$E(SSTr) = \sum_{i=1}^m \sigma^2 + \left[\sum_{i=1}^m n_i \mu_i^2 \right] - \sigma^2 - n \bar{\mu}^2$$

$$E(SSTr) = (m-1)\sigma^2 + \left[\sum_{i=1}^m n_i (\mu_i^2 - \bar{\mu}^2) \right] = (m-1)\sigma^2 + [\sum_{i=1}^m n_i (\mu_i - \bar{\mu})^2]$$

$$E[MStr] = E \left(\frac{SSTr}{m-1} \right) = \sigma^2 + \frac{1}{m-1} [\sum_{i=1}^m n_i (\mu_i - \bar{\mu})^2] = \sigma^2 \text{ (if the null hypothesis is true)}$$

because $[\sum_{i=1}^m n_i (\mu_i - \bar{\mu})^2] = 0$, if the null hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_m = \bar{\mu}$ is true.

So $MStr$ is an unbiased estimator of σ^2 if the null hypothesis is true.

$E(MStr) = \sigma^2$ if H_0 is true

For any random variable $E[Y^2] = V(Y) + (E[Y])^2$

$$\begin{aligned} E(SStr) &= E\left[\frac{1}{J}\sum_i X_{i.}^2 - \frac{1}{IJ}X_{..}^2\right] = \frac{1}{J}\sum_i E(X_{i.}^2) - \frac{1}{IJ}E(X_{..}^2) \quad ([E(aX + bY) = aE(X) + bE(Y)]) \\ &= \frac{1}{J}\sum_i\{V(X_{i.}) + [E(X_{i.})]^2\} - \frac{1}{IJ}\{V(X_{..}) + [E(X_{..})]^2\} \quad (X_{i.} = \sum_{j=1}^J X_{ij} \quad X_{..} = \sum_{i=1}^I \sum_{j=1}^J X_{ij}) \end{aligned}$$

Since $E(X_{..}) = E(\sum_{i=1}^I \sum_{j=1}^J X_{ij}) = E(\sum_{i=1}^I X_{i.}) = \sum_i E(X_{i.}) = \sum_i J\mu_i = J\sum_i \mu_i = IJ\bar{\mu}$, where $\bar{\mu} = \frac{\sum_i \mu_i}{I}$

$$\begin{aligned} E(SStr) &= \frac{1}{J}\sum_i\{J\sigma^2 + [J\mu_i]^2\} - \frac{1}{IJ}[IJ\sigma^2 + [IJ\bar{\mu}]^2] \quad (X_{i.} = X_{i1} + X_{i2} + \dots + X_{iJ}) \\ &= \frac{1}{J}\sum_i\{J\sigma^2 + J^2\mu_i^2\} - \frac{1}{IJ}[IJ\sigma^2 + [IJ\bar{\mu}]^2] \\ &= I\sigma^2 + \sum_i\{J\mu_i^2\} - \sigma^2 - IJ\bar{\mu}^2 \\ &= (I-1)\sigma^2 + \sum_i\{J\mu_i^2\} - IJ\bar{\mu}^2 = (I-1)\sigma^2 + J(\sum_i \mu_i^2 - I\bar{\mu}^2) \\ &= (I-1)\sigma^2 + J\sum_i(\mu_i - \bar{\mu})^2 \\ (\sum_i(\mu_i - \bar{\mu})^2) &= \sum_i \mu_i^2 - 2\bar{\mu} \sum_i \mu_i + I\bar{\mu}^2 = \sum_i \mu_i^2 - 2I\bar{\mu}^2 + I\bar{\mu}^2 = \sum_i \mu_i^2 - I\bar{\mu}^2 \end{aligned}$$

$$E(MStr) = E\left(\frac{SStr}{I-1}\right) = \sigma^2 + \frac{J}{I-1}\sum_i(\mu_i - \bar{\mu})^2$$

If H_0 is true, $\mu_1 = \mu_2 = \dots = \mu_I = \bar{\mu}$ and $\sum_i(\mu_i - \bar{\mu})^2 = 0$. So $E(MStr) = \sigma^2$.

Theoretical Results | Stat 414/415 – Statistics Online – Penn State

<https://online.stat.psu.edu/stat415/lesson/13/13.3>

Number of population : m , the sample size of i^{th} population : n_i

$$SSTr = \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^m n_i \bar{X}_{i.}^2 - n \bar{X}_{..}^2$$

$$E(SSTr) = E \left[\sum_{i=1}^m n_i \bar{X}_{i.}^2 - n \bar{X}_{..}^2 \right] = \left[\sum_{i=1}^m n_i E(\bar{X}_{i.}^2) \right] - n E \left[\bar{X}_{..}^2 \right]$$

Since $E[X^2] = Var(X) + \mu^2$,

$$E(SSTr) = \left[\sum_{i=1}^m n_i \left(\frac{\sigma^2}{n_i} + \mu_i^2 \right) \right] - n \left[\frac{\sigma^2}{n} + \bar{\mu}^2 \right]$$

where $\bar{\mu} = \sum_{i=1}^m n_i \mu_i / n$ and $E(\bar{X}_{..}) = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} E(X_{ij}) = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \mu_i = \frac{1}{n} \sum_{i=1}^m n_i \mu_i = \bar{\mu}$

$$E(SSTr) = \sum_{i=1}^m \sigma^2 + \left[\sum_{i=1}^m n_i \mu_i^2 \right] - \sigma^2 - n \bar{\mu}^2$$

$$E(SSTr) = (m - 1)\sigma^2 + \left[\sum_{i=1}^m n_i (\mu_i^2 - \bar{\mu}^2) \right] = (m - 1)\sigma^2 + [\sum_{i=1}^m n_i (\mu_i - \bar{\mu})^2]$$

$$E[MStr] = E \left(\frac{SSTr}{m-1} \right) = \sigma^2 + \frac{1}{m-1} [\sum_{i=1}^m n_i (\mu_i - \bar{\mu})^2] = \sigma^2 \text{ (if the null hypothesis is true)}$$

because $[\sum_{i=1}^m n_i (\mu_i - \bar{\mu})^2] = 0$, if the null hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_m = \bar{\mu}$ is true.

So $MStr$ is an unbiased estimator of σ^2 if the null hypothesis is true.

F Distribution and the F Test

- Let $F = MStr / MSE$ be the test statistic in a single-factor ANOVA problem involving I populations or treatments with a random sample of J observations from each one.
- When H_0 is true, F has an F distribution with $\nu_1 = I - 1$ and $\nu_2 = I(J - 1)$.
- With f denoting the computed value of F , the rejection region $f \geq F_{\alpha, I-1, I(J-1)}$ specifies a test with significance level α .
- If the null hypothesis is false, we have $MStr > MSE$. Therefore we have single rejection region.
- <https://online.stat.psu.edu/stat415/lesson/13/13.3>
- SSE/σ^2 follows chi-square distribution with degrees of freedom $IJ - I = I(J - 1)$.
- When H_0 is true, $SSTr/\sigma^2$ follows chi-square distribution with degrees of freedom $I - 1$.

$$F = \frac{\frac{SSTr}{\sigma^2}/(I-1)}{\frac{SSE}{\sigma^2}/(I(J-1))} = \frac{MStr}{MSE} \sim F_{I-1, I(J-1)}$$

10.1 Single Factor ANOVA Example 10.1

We want to compare 4 types of boxes with respect to the compression strength.

We obtain 6 observations for each type of box.

Type of Box	Compression Strength (lb)							Sample Mean	Sample SD
1	655.5	788.3	734.3	721.4	679.1	699.4		713.00	46.55
2	789.2	772.5	786.9	686.1	732.1	774.8		756.93	40.34
3	737.1	639.0	696.3	671.7	717.2	727.1		698.07	37.20
4	535.1	628.7	542.4	559.0	586.9	520.0		562.02	39.87
					Grand mean =		682.50		

We want to test if all four types have the same mean compression strength.

How to construct the appropriate hypotheses? How to perform the test?

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a: \text{at least two of the } \mu_i \text{'s are different}$$

Example 10.2 (Example 10.1 continued) :

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs $H_1 : \text{At least two } \mu_i\text{'s are different}$

$$\bar{x}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^J x_{ij}}{IJ} = 682.50$$

$$MStr = \frac{6}{4-1} [(713.00 - 682.50)^2 + (756.93 - 682.50)^2 + (698.07 - 682.50)^2 + (562.02 - 682.50)^2] = 42455.86$$

$$MSE = \frac{1}{4} [(46.55)^2 + (40.34)^2 + (37.20)^2 + (39.87)^2] = 1691.92$$

$$f = MStr / MSE = \frac{42455.86}{1691.92} = 25.09$$

Since $25.09 \geq 3.10 = F_{0.05,3,20}$, H_0 is rejected at significance level 0.05.

$$\nu_1 = I - 1 = 4 - 1 = 3, \nu_2 = I(J - 1) = 4(6 - 1) = 20$$

> qf(0.95, 3, 20)

[1] 3.098391

R execution of Example 10.1

```
> compression <- c(655.5,788.3,734.3,721.4,679.1,699.4,  
+ 789.2,772.5,786.9,686.1,732.1,774.8,  
+ 737.1,639.0,696.3,671.7,717.2,727.1,  
+ 535.1,628.7,542.4,559.0,586.9,520.0)  
> type <- c(rep("1", 6), rep("2", 6), rep("3", 6), rep("4", 6))  
> oneway <- aov(compression ~ type)  
> summary(oneway)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	127375	42458	25.09	5.53e-07 ***
Residuals	20	33839	1692		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Sums of Squares

The total sum of squares (SST), treatment sum of squares (SSTr), and error sum of squares (SSE) are given by

$$SST = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2 - \frac{1}{IJ} \bar{x}_{..}^2$$

$$SSTr = \sum_{i=1}^I \sum_{j=1}^J (\bar{x}_{i.} - \bar{x}_{..})^2 = \frac{1}{J} \sum_{i=1}^I \bar{x}_{i.}^2 - \frac{1}{IJ} \bar{x}_{..}^2$$

$$SSE = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \bar{x}_{i.})^2$$

Where $x_{i.} = \sum_{j=1}^J x_{ij}$, $\bar{x}_{..} = \sum_{i=1}^I \sum_{j=1}^J x_{ij}$

Fundamental Identity : $SST = SSTr + SSE$

Sum of Squares

$$\begin{aligned} SST &= \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^I \sum_{j=1}^J (x_{ij}^2 - 2\bar{x}_{..}x_{ij} + \bar{x}_{..}^2) \\ &= \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2 - 2\bar{x}_{..} \sum_{i=1}^I \sum_{j=1}^J x_{ij} + IJ\bar{x}_{..}^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2 - 2IJ\bar{x}_{..}^2 + IJ\bar{x}_{..}^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2 - IJ\bar{x}_{..}^2 = \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2 - \frac{1}{IJ} x_{..}^2 \end{aligned}$$

$$\begin{aligned} SSTR &= \sum_{i=1}^I \sum_{j=1}^J (\bar{x}_{i.} - \bar{x}_{..})^2 = J \sum_{i=1}^I (\bar{x}_{i.} - \bar{x}_{..})^2 \\ &= J \sum_{i=1}^I (\bar{x}_{i.}^2 - 2\bar{x}_{..}\bar{x}_{i.} + \bar{x}_{..}^2) = J \left\{ \sum_{i=1}^I \left(\frac{\bar{x}_{i.}}{J} \right)^2 - 2 \frac{\bar{x}_{..}}{IJ} \sum_{i=1}^I \frac{\bar{x}_{i.}}{J} + I \left(\frac{\bar{x}_{..}}{IJ} \right)^2 \right\} \\ &= J \left\{ \sum_{i=1}^I \left(\frac{\bar{x}_{i.}}{J} \right)^2 - 2I \frac{\bar{x}_{..}}{IJ} \frac{\bar{x}_{..}}{IJ} + I \left(\frac{\bar{x}_{..}}{IJ} \right)^2 \right\} = \frac{1}{J} \sum_i x_{i.}^2 - \frac{1}{IJ} x_{..}^2 \\ &= J \left\{ \sum_{i=1}^I \left(\frac{\bar{x}_{i.}}{J} \right)^2 - I \left(\frac{\bar{x}_{..}}{IJ} \right)^2 \right\} = \frac{1}{J} \sum_i x_{i.}^2 - \frac{1}{IJ} x_{..}^2 \end{aligned}$$

$$MStr = \frac{SSTr}{I - 1}, MSE = \frac{SSE}{I(J - 1)}, F = \frac{MStr}{MSE}$$

Table 10.2 An ANOVA Table

Source of Variation	df	Sum of Squares	Mean Square	f
Treatments	I-1	$SSTr$	$MStr$	$MStr/ MSE$
Error	$I(J-1)$	SSE	MSE	
Total	$IJ-1$	SST		

Example 10.4

The accompanying data shows the degree of soiling for fabric copolymerized with three different mixtures of methacrylic acid.

		x_{ij}	\bar{x}_{ij}	sd
Mixture 1	0.56	1.12	0.90	1.07
Mixture 2	0.94			
Mixture 3	0.72	0.69	0.87	0.78
	0.91			
	0.62	1.08	1.07	0.99
	0.93			
		4.59	0.918	0.220
		3.97	0.794	0.094
		4.69	0.938	0.188

$$SST = \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2 - \frac{1}{IJ} \bar{x}_{..}^2 = (0.56)^2 + (1.12)^2 + \dots + (0.93)^2 - \frac{(13.25)^2}{15} = 0.4309$$

$$SSTR = \frac{1}{J} \sum_{i=1}^I \bar{x}_{ij}^2 - \frac{1}{IJ} \bar{x}_{..}^2 = \frac{1}{5} \{(4.59)^2 + (3.97)^2 + (4.69)^2\} - \frac{(13.25)^2}{15} = 0.0608$$

$$SSE = 0.4309 - 0.0608 = 0.3701$$

$$SST = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \bar{x}_{..})^2 = (0.56 - 0.883)^2 + (1.12 - 0.883)^2 + \dots + (0.93 - 0.883)^2 = 0.4309$$

$$SSTR = \sum_{i=1}^I \sum_{j=1}^J (\bar{x}_{ij} - \bar{x}_{..})^2 = 5 \{(0.918 - 0.883)^2 + (0.794 - 0.883)^2 + (0.938 - 0.883)^2\} = 0.060855$$

$$SSE = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \bar{x}_{ij})^2 = (5 - 1) \times [(0.22)^2 + (0.094)^2 + (0.188)^2] = 0.3701$$

$$s = \sqrt{\frac{1}{I-1} \sum_{i=1}^I (\bar{x}_{ij} - \bar{x}_{..})^2 / (I-1)} = \sqrt{\frac{1}{4} \sum_{i=1}^I ((0.918 - 0.883)^2 + (0.794 - 0.883)^2 + (0.938 - 0.883)^2) / 4} = 0.188$$

```
> degree <- c(0.56, 1.12, 0.90, 1.07, 0.94,  
+ 0.72, 0.69, 0.87, 0.78, 0.91,  
+ 0.62, 1.08, 1.07, 0.99, 0.93)  
> sst <- sum(degree^2) - sum(degree)^2/15; sst  
[1] 0.4309333  
> sstr <- (4.59^2+3.97^2+4.69^2)/5 - 13.25^2/15; sstr  
[1] 0.06085333  
> sse <- sst-sstr; sse  
[1] 0.37008  
> sst1 <- sum((degree - mean(degree))^2); sst1  
[1] 0.4309333
```

Example 10.4

```
> degree <- c(0.56, 1.12, 0.90, 1.07, 0.94,  
+ 0.72, 0.69, 0.87, 0.78, 0.91,  
+ 0.62, 1.08, 1.07, 0.99, 0.93)  
> type <- c(rep("1", 5), rep("2", 5), rep("3", 5))  
> oneway <- aov(degree ~ type)  
> summary(oneway)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	2	0.0609	0.03043	0.987	0.401
Residuals	12	0.3701	0.03084		
<hr/>					
Source of Variation	df	Sum of Squares	Mean Square	f	
Treatments	2	0.0608	0.0304	0.987	
Error	12	0.3701	0.0308		
Total	14	0.4309			

General ANOVA Procedure

1. Data Layout

Treatment	Quantity of interest	Group Sample Mean	Group Sample Variance
1	$X_{11}, X_{12}, \dots, X_{1J}$	$\bar{X}_{1..}$	S_1^2
2	$X_{21}, X_{22}, \dots, X_{2J}$	$\bar{X}_{2..}$	S_2^2
:	\vdots	\vdots	\vdots
I	$X_{I1}, X_{I2}, \dots, X_{IJ}$	$\bar{X}_{I..}$	S_I^2
	Grand mean	$\bar{X}_{...}$	

2. Assumptions

$X_{11}, X_{12}, \dots, X_{1J}$ are from $N(\mu_1, \sigma^2)$

$X_{21}, X_{22}, \dots, X_{2J}$ are from $N(\mu_2, \sigma^2)$

\vdots

$X_{I1}, X_{I2}, \dots, X_{IJ}$ are from $N(\mu_I, \sigma^2)$

General ANOVA Procedure

3. Hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I \quad \text{vs} \quad H_a : \text{At least two group means are different}$$

4. Mean square for treatment

$$MSTr = \frac{J}{I-1} \sum_i (\bar{X}_{i\cdot} - \bar{X}_{..})^2$$

5. Mean square for error

$$MSE = \frac{s_1^2 + s_2^2 + \dots + s_I^2}{I}$$

6. Proposition

When H_0 is true, $E(MSTr) = E(MSE) = \sigma^2$

When H_0 is false, $E(MSTr) > E(MSE) = \sigma^2$

7. Test Statistic

$$F = \frac{MSTr}{MSE} \sim F_{I-1, I(J-1)}$$

General ANOVA Procedure

8. Rejection region :

$$f > F_{\alpha, I-1, J(I-1)}$$

Source of Variation	df	Sum of Squares	Mean Square	f
Treatments	I-1	$SSTr$	$MStr$	$MStr / MSE$
Error	I(J-1)	SSE	MSE	
Total	IJ-1	SST		

Example 10.4 An ANOVA Table

Source of Variation	df	Sum of Squares	Mean Square	f
Treatments	2	0.0608	0.0304	0.99
Error	12	0.3701	0.0308	
Total	14	0.4309		

Because $f=0.99 < F_{0.01,2,12} = 6.93$, H_0 is not rejected at significance level 0.01.

10.2 Multiple Comparisons in ANOVA

In this section we will learn how to estimate the difference between any pairs of means.

- **Tukey's Procedure**

Tukey's procedure involves the use of another probability distribution called the Studentized range distribution.

The distribution depends on two parameters: a numerator df m and a denominator df ν . Denote $Q_{\alpha,m,\nu}$ to be the upper α cut-off value (see Appendix Table A.10).

For a specific confidence level $1 - \alpha$ the **simultaneous confidence intervals** for the differences are

given by

$$(\bar{X}_i - \bar{X}_j) + Q_{\alpha, L(L-1)} \sqrt{\frac{MSE}{n}}$$

- Each interval that does not include 0 yields the conclusion that the corresponding values of μ_i and μ_j differ significantly from one another.
- An easy way to detect such differences without computing the actual interval is as follows:
 - First select α and determine the value of $Q_{\alpha,I,I(J-1)}$ (Table A.10)
 - Obtain MSE from the ANOVA table
 - Compute $w = Q_{\alpha,I,I(J-1)} \sqrt{\frac{MSE}{J}}$
 - List the sample means in *increasing order* and underline those pairs that differ by less than w
 - Any pair of sample means not underlined by the same line corresponds to a pair of population or treatment means that are judged significantly different.

Example 10.5

An experiment was carried out to compare five different brands of automobile oil filters with respect to their ability to capture foreign material.

Let μ_i denote the true average amount of material captured by brand i filters ($i = 1, \dots, 5$) under controlled conditions.

A sample of nine filters of each brand was used, resulting in the following sample mean amounts: $\bar{x}_1 = 14.5$, $\bar{x}_2 = 13.8$, $\bar{x}_3 = 13.3$, $\bar{x}_4 = 14.3$, and $\bar{x}_5 = 13.1$.

The MSE from ANOVA is $MSE = 0.088$.

The test statistic value was $f = 37.84$, which is found significant at $\alpha = 0.05$ level.
Find out which means are different.

Hint: From Appendix Table A.10 $Q_{0.05,5,40} = 4.04$.

```
> qtukey(0.95, 5, 40)
```

```
[1] 4.039123
```

Source of Variation	df	Sum of Squares	Mean Square	f
Treatments	4	13.32	3.33	37.84
Error	40	3.53	0.088	
Total	44	16.85		

Because $f=37.84 > F_{0.05,4,40} = 2.61$, H_0 is rejected at significance level 0.05.

$$\text{Since } Q_{0.05,5,40} = 4.04, w = 4.04\sqrt{0.088/9} = 0.4$$

Arrange the five sample means in increasing order and underline those pairs that differ by less than w .

$$\begin{array}{ccccc} \bar{x}_5. & \bar{x}_3. & \bar{x}_2. & \bar{x}_4. & \bar{x}_1. \\ \underline{13.1} & \underline{13.3} & 13.8 & \underline{14.3} & \underline{14.5} \end{array}$$

Thus brands 1 and 4 are not significantly different from one another, but are significantly higher than the other three brands. Brand 2 is significantly better than 3 and 5 but worse than 1 and 4. Brands 3 and 5 do not differ significantly.

Example 10.4

The accompanying data shows the degree of soiling for fabric copolymerized with three different mixtures of methacrylic acid.

				x_i	\bar{x}_i	sd	
Mixture 1	0.56	1.12	0.90	1.07	4.59	0.918	0.220
Mixture 2	0.94				3.97	0.794	0.094
Mixture 3	0.72	0.69	0.87	0.78	4.69	0.938	0.188
	0.91						
	0.62	1.08	1.07	0.99			
	0.93						
					$x_{..} = 13.25$	$\bar{x}_{..} = 0.883$	

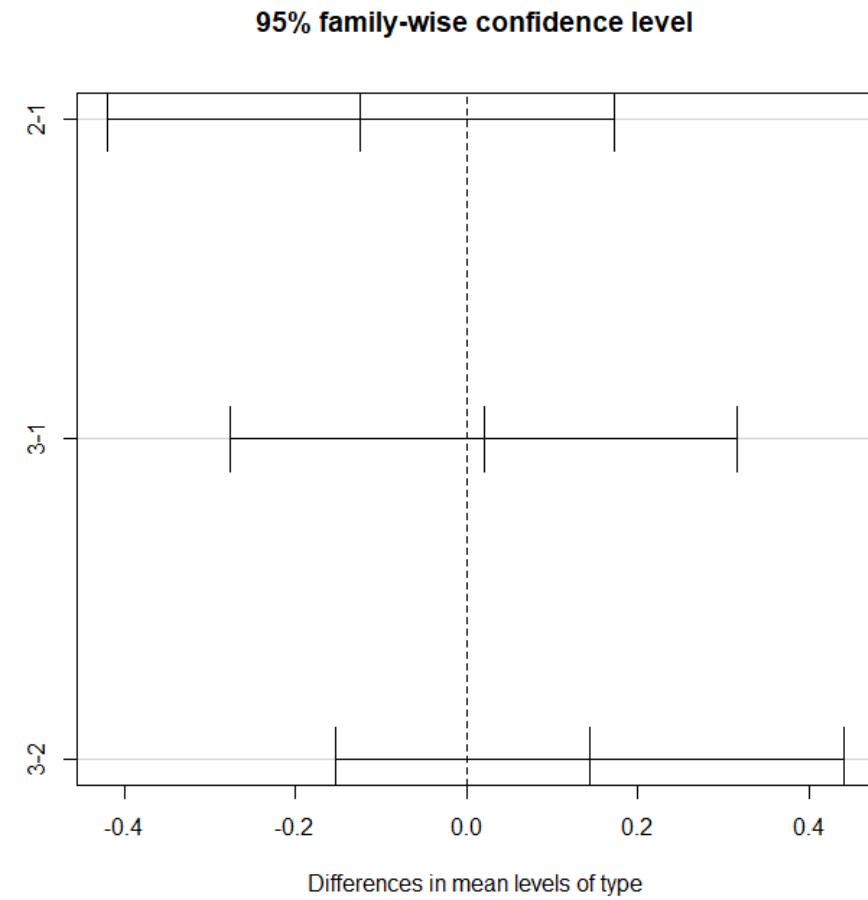
$H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a : \text{At least two group means are different}$

Tukey's Comparison : Example 10.4

```
> degree <- c(0.56, 1.12, 0.90, 1.07, 0.94,  
+ 0.72, 0.69, 0.87, 0.78, 0.91,  
+ 0.62, 1.08, 1.07, 0.99, 0.93)  
> type <- c(rep("1", 5), rep("2", 5), rep("3", 5))  
> oneway <- aov(degree ~ type)  
> summary(oneway)  
> posthoc <- TukeyHSD(x=oneway, 'type', conf.level=0.95)  
> posthoc
```

	diff	lwr	upr	p adj
2-1	-0.124	-0.4203131	0.1723131	0.5225991
3-1	0.020	-0.2763131	0.3163131	0.9823093
3-2	0.144	-0.1523131	0.4403131	0.4235977

```
> plot(posthoc)
```



Source of Variation	df	Sum of Squares	Mean Square	f
type	2	0.0609	0.03043	0.987
Error	12	0.3701	0.03084	
Total	14	0.4310		

Because $f=0.987 < F_{0.05,2,12} = 3.106$, H_0 is not rejected at significance level 0.05.

$$\text{Since } Q_{0.05,3,12} = 3.77, w = 3.77\sqrt{0.03084/5} = 0.296$$

$$2-1 : (\bar{X}_{2.} - \bar{X}_{1.}) \pm Q_{0.05,3,12} \sqrt{\frac{MSE}{J}} = (0.794 - 0.918) \pm 0.296 \Rightarrow (-0.42, 0.172)$$

$$3-1 : (\bar{X}_{3.} - \bar{X}_{1.}) \pm Q_{0.05,3,12} \sqrt{\frac{MSE}{J}} = (0.938 - 0.918) \pm 0.296 \Rightarrow (-0.276, 0.316)$$

$$3-2 : (\bar{X}_{3.} - \bar{X}_{2.}) \pm Q_{0.05,3,12} \sqrt{\frac{MSE}{J}} = (0.938 - 0.794) \pm 0.296 \Rightarrow (-0.152, 0.44)$$

Unequal Sample Sizes

J_1, J_2, \dots, J_I : the size of I samples

$n = \sum_i J_i$: the total number of observation

$$SST = \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^I \sum_{j=1}^{J_i} X_{ij}^2 - \frac{1}{n} \bar{X}_{..}^2 \quad df = n - 1$$

$$SSTR = \sum_{i=1}^I \sum_{j=1}^{J_i} (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^I \frac{1}{J_i} \bar{X}_{i.}^2 - \frac{1}{n} \bar{X}_{..}^2 \quad df = I - 1$$

$$SSE = \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{i.})^2 = SST - SSTR \quad df = \sum(J_i - 1) = n - I$$

Test statistic value :

$$f = \frac{MSTR}{MSE} \quad \text{where } MSTR = \frac{SSTR}{I-1} \quad \text{and } MSE = \frac{SSE}{n-I}$$

Rejection region : $f > F_{\alpha, I-1, n-I}$

Example 10.9

The following data shows the elastic modulus obtained for specimens of a certain alloy produced by three different casting processes.

		J_i	$x_{i.}$	$\bar{x}_{i.}$
Permanent molding	45.5 45.3 45.4 44.4 44.6 43.9 44.6 44.0	8	357.7	44.712
Die casting	44.2 43.9 44.7 44.2 44.0 43.8 44.6 43.1	6	273.5	44.062
Plaster molding	46.0 45.9 44.8 46.2 45.1 45.5	22	983.7	45.583
		$\bar{x}_{..} = 983.7$	$\bar{x}_{..} = \frac{3}{3} = 44.713$	

$H_0: \mu_1 = \mu_2 = \mu_3$ vs $H_a:$ at least two of the μ_i 's are different

$$SST = \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{..})^2 = (45.5 - 44.713)^2 + (45.3 - 44.713)^2 + \dots + (45.5 - 44.713)^2 = 13.926$$

$$\begin{aligned} SSTr &= \sum_{i=1}^I \sum_{j=1}^{J_i} (\bar{X}_{i.} - \bar{X}_{..})^2 = \{8(44.7125 - 44.713)^2 + 8(44.0625 - 44.713)^2 + 6(45.58335 - 44.713)^2\} \\ &= 7.93 \end{aligned}$$

Example 10.9

```
> modulus <- c(45.5, 45.3, 45.4, 44.4, 44.6, 43.9, 44.6, 44.0,  
+ 44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1,  
+ 46.0, 45.9, 44.8, 46.2, 45.1, 45.5)  
> type <- c(rep("1", 8), rep("2", 8), rep("3", 6))  
> oneway <- aov(modulus ~ type)  
> summary(oneway)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	2	7.930	3.965	12.56	0.000334 ***
Residuals	19	5.996	0.316		
Source of Variation					
	df	Sum of Squares	Mean Square	f	
Treatments	2	7.93	3.965	12.56	
Error	19	5.996	0.316		
Total	21	13.93			

Tukey's Comparison : Example 10.9

```
> posthoc <- TukeyHSD(x=oneway, 'type', conf.level=0.95)
> posthoc
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = modulus ~ type)

	diff	lwr	upr	p adj
2-1	-0.6500000	-1.3635569	0.06355694	0.0780174
3-1	0.8708333	0.1001038	1.64156292	0.0253295
3-2	1.5208333	0.7501038	2.29156292	0.0002183

F Distribution and the F Test

$$\sum \sum x_i^2 = 43998.73 \quad CF = \frac{983.7^2}{22} = 43984.8$$

$$SST = 43998.73 - 43984.8 = 13.93$$

$$SSTr = \frac{(357.7)^2}{8} + \frac{(352.5)^2}{8} + \frac{(273.5)^2}{6} - 43984.8 = 7.93$$

$$SSE = 13.93 - 7.93 = 6.00$$

Source of variation	df	Sum of squares	Mean square	f
Treatment	2	7.93	3.965	12.56
Error	19	6.00	0.3158	
Total	21	13.93		

Since $12.56 \geq F_{0.001,2,19} = 10.16$, H_0 is rejected at significance level 0.001.

$$\text{Let } w_{ij} = Q_{\alpha, I, n-I} \sqrt{\frac{MSE}{2} \left(\frac{1}{J_i} + \frac{1}{J_j} \right)}$$

Then the probability is $1 - \alpha$ that

$$(\bar{X}_{i\cdot} - \bar{X}_{j\cdot}) - w_{ij} \leq \mu_i - \mu_j \leq (\bar{X}_{i\cdot} - \bar{X}_{j\cdot}) + w_{ij} \quad (\bar{X}_{i\cdot} - \bar{X}_{j\cdot}) \pm w_{ij}$$

Example 10.10

In Example 10.9, $J_1 = 8, J_2 = 8, J_3 = 6, I = 3, n - I = 19, MSE = 0.316, Q_{0.05, 3, 19} = 3.59$

$$w_{12} = 3.59 \sqrt{\frac{0.316}{2} \left(\frac{1}{8} + \frac{1}{8} \right)} = 0.713, w_{13} = 0.771, w_{23} = 0.771$$

Since $\bar{x}_{1\cdot} - \bar{x}_{2\cdot} = 44.71 - 44.06 = 0.65 < w_{12}$, μ_1 and μ_2 are not significantly different.

$$2-1 : -0.65 \pm 0.713 \rightarrow (-1.363, 0.063)$$

$$3-1 : 0.87 \pm 0.771 \rightarrow (0.099, 1.641)$$

12.3 Inferences About the Slope Parameter β_1

The values of x_i 's are assumed to be chosen before the experiment is performed, so only the Y_i 's are random.

The estimators for β_0 and β_1 are obtained by replacing y_i by Y_i .

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n}$$

The denominator of $\hat{\beta}_1$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, depends only on the x_i 's and not on the Y_i 's, so it is a constant.

Because $\sum_{i=1}^n (x_i - \bar{x}) \bar{Y} = \bar{Y} \sum_{i=1}^n (x_i - \bar{x}) = \bar{Y} \cdot 0 = 0$, the slope estimator can be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}} = \sum_{i=1}^n c_i Y_i, \text{ where } c_i = (x_i - \bar{x}) / S_{xx}$$

12.3 Inferences About the Slope Parameter β_1

Proposition

1. The mean value of $\hat{\beta}_1$: $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 .
2. The variance and standard deviation of $\hat{\beta}_1$:

$$V(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}}, \quad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

Replacing σ by its estimate s gives an estimate of $\sigma_{\hat{\beta}_1}$.

3. The estimator $\hat{\beta}_1$ has a normal distribution (because it is a linear function of independent normal rv's).

cf)<https://stats.stackexchange.com/questions/12186/expected-value-and-variance-of-estimation-of-slope-parameter-beta-1-in-simple>

Theorem

The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

has a t distribution with $(n - 2)$ df.

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}\right) = E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{S_{xx}}\right) \\
&= E\left(\frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right) \\
&= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i}{S_{xx}}\right) = \beta_1 E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{S_{xx}}\right) = \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{S_{xx}} = \beta_1
\end{aligned}$$

Here $\sum_{i=1}^n (x_i - \bar{x}) x_i = \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) \bar{x} = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = S_{xx}$

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}\right) = Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{S_{xx}}\right) \\
&= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_0}{S_{xx}} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \beta_1 x_i}{S_{xx}} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right) \\
&= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right) = Var\left(\frac{(x_1 - \bar{x}) \varepsilon_1}{S_{xx}} + \dots + \frac{(x_n - \bar{x}) \varepsilon_n}{S_{xx}}\right) \\
&= \frac{(x_1 - \bar{x})^2 \sigma^2}{(S_{xx})^2} + \dots + \frac{(x_n - \bar{x})^2 \sigma^2}{(S_{xx})^2} = \sigma^2 \left[\frac{\sum (x_i - \bar{x})^2}{(S_{xx})^2} \right] = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

A Confidence Interval for β_1

We begin with a probability statement

$$P\left(-t_{\frac{\alpha}{2}, n-2} < \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} < t_{\frac{\alpha}{2}, n-2}\right) = 1 - \alpha$$

A $100(1-\alpha)\%$ CI for the slope β_1 of the true regression line is

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{\beta}_1}$$

$$-t_{\frac{\alpha}{2}, n-2} < \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} < t_{\frac{\alpha}{2}, n-2}$$

$$-t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} < \hat{\beta}_1 - \beta_1 < t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$$

$$-\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} < -\beta_1 < -\hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$$

$$\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}$$

Example

The data on the next table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels. Compute the 95% CI for the slope β_1 .

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4
	54.6	58.8	58.0								

$$\begin{aligned}\hat{\beta}_1 \pm t_{0.025,12} \cdot s_{\hat{\beta}_1} &= \hat{\beta}_1 \pm t_{0.025,12} \cdot \frac{s}{\sqrt{s_{xx}}} \\ &= -0.20939 \pm 2.179 \times \frac{2.564}{82.4789} \\ &= -0.20939 \pm 2.179 \times 0.03109 \rightarrow (-0.2771, 0.1416)\end{aligned}$$

Steps in the analysis

```
> f <- lm(y ~ x)  
> summary(f)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.9488	-1.5665	0.6817	1.0846	4.8974

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	75.21243	2.98363	25.208	9.22e-12 ***
x	-0.20939	0.03109	-6.734	2.09e-05 ***

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	'.'	0.1 ''	1	

Residual standard error: 2.564 on 12 degrees of freedom

Multiple R-squared: 0.7908, Adjusted R-squared: 0.7733

F-statistic: 45.35 on 1 and 12 DF, p-value: 2.091e-05

Hypothesis-Testing Procedures

Null hypothesis : $H_0: \beta_1 = \beta_{10}$

Test statistic value : $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Alternative Hypothesis

$$H_a: \beta_1 > \beta_{10}$$

$$H_a: \beta_1 < \beta_{10}$$

$$H_a: \beta_1 \neq \beta_{10}$$

Rejection Region for Level α Test

$$t \geq t_{\alpha, n-2}$$

$$t \leq -t_{\alpha, n-2}$$

$$\text{either } t \geq t_{\alpha/2, n-2} \text{ or } t \leq -t_{\alpha/2, n-2}$$

A P -value based on $n - 2$ df can be calculated just as was done previously.

The model utility test is the test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, in which case the test statistic value is the t ratio of $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.

Example

The data on the next table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels. Find the P -value for $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4
	54.6	58.8	58.0								

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{-0.20939}{0.03109} = -6.734$$

$$P\text{-value} = P(t_{14-2} < -6.734) + P(t_{14-2} > 6.734) = 2 \times P(t_{14-2} < -6.734) = 2.091352e-05$$

This value is the same as the value at the column of $\Pr(>|t|)$ on page 10.

```
> 2*pt(-6.734, 12)
```

```
[1] 2.091352e-05
```

Since $t = -6.734 \leq -t_{0.025, 12} = -2.179$, H_0 can be rejected under the level 0.05 test.

Inferences Concerning $\mu_{Y|x^*}$ and the Prediction of Future Y Values

Proposition

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is some fixed value of x . Then

1. The mean value of \hat{Y} is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

2. The variance of \hat{Y} is

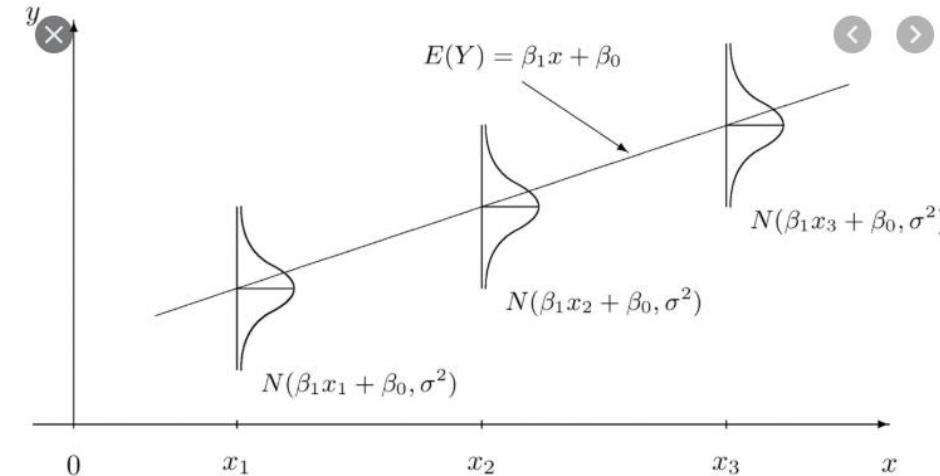
$$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

3. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ results from replacing σ by its estimate s .

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

A $100(1-\alpha)\%$ CI for $\mu_{Y|x^*}$, the expected value of Y when $x = x^*$, is $\frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{y}}$$



$$V(\hat{Y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) = Var(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x^*) = Var(\bar{y} + \hat{\beta}_1(x^* - \bar{x}))$$

$$= Var(\bar{y}) + (x^* - \bar{x})^2 Var(\hat{\beta}_1) + 2(x^* - \bar{x}) Cov(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2(x^* - \bar{x})^2}{S_{xx}} + 0$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{\sum y_i}{n}, \sum c_i y_i\right) \quad (c_i = (x_i - \bar{x})/S_{xx})$$

$$= \frac{1}{n} \sum c_i Var(y_i)$$

$$= \frac{\sigma^2}{n} \sum c_i = 0$$

> <https://math.stackexchange.com/questions/2912624/show-that-cov-bary-hat-beta-1-0>

A Prediction Interval of Future Y Values

Consider some future observation value Y when the independent variable has value x^* .

The error of prediction is $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, a difference between two random variables.

Because the future value Y is independent of the observed Y_i 's,

$$\begin{aligned} V(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)) &= V(Y) + V(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] \end{aligned}$$

A $100(1-\alpha)\%$ CI for a future observation Y to be made when $x = x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Example

The data on the table is x = iodine value (g) and y = cetane number for a sample of 14 biofuels.

- Compute the 95% CI for $\mu_{Y|100}$, the expected value of Y when $x = 100$

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0
	81.5	71.0	69.2								
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	
	61.4	54.6	58.8	58.0							

$$\hat{\beta}_0 + \hat{\beta}_1 100 \pm t_{0.025,12} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{\beta}_0 + \hat{\beta}_1 100 \pm t_{0.025,12} \cdot s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}}$$
$$= 75.21243 + (-0.20939) \times 100 \pm 2.179 \cdot 2.564 \sqrt{\frac{1}{14} + \frac{(100 - 93.393)^2}{6802.769}} \rightarrow (52.715, 55.832)$$

```
> predict(f, list(x=c(100)), interval="conf")
```

- Compute the 95% CI for a future observation Y when $x = 100$ is

$$\hat{\beta}_0 + \hat{\beta}_1 100 \pm t_{0.025,12} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{\beta}_0 + \hat{\beta}_1 100 \pm t_{0.025,12} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}}$$
$$= 75.21243 + (-0.20939) \times 100 \pm 2.179 \cdot 2.564 \sqrt{1 + \frac{1}{14} + \frac{(100 - 93.393)^2}{6802.769}} \rightarrow (48.473, 60.074)$$

```
> predict(f, list(x=c(100)), interval="pred")
```

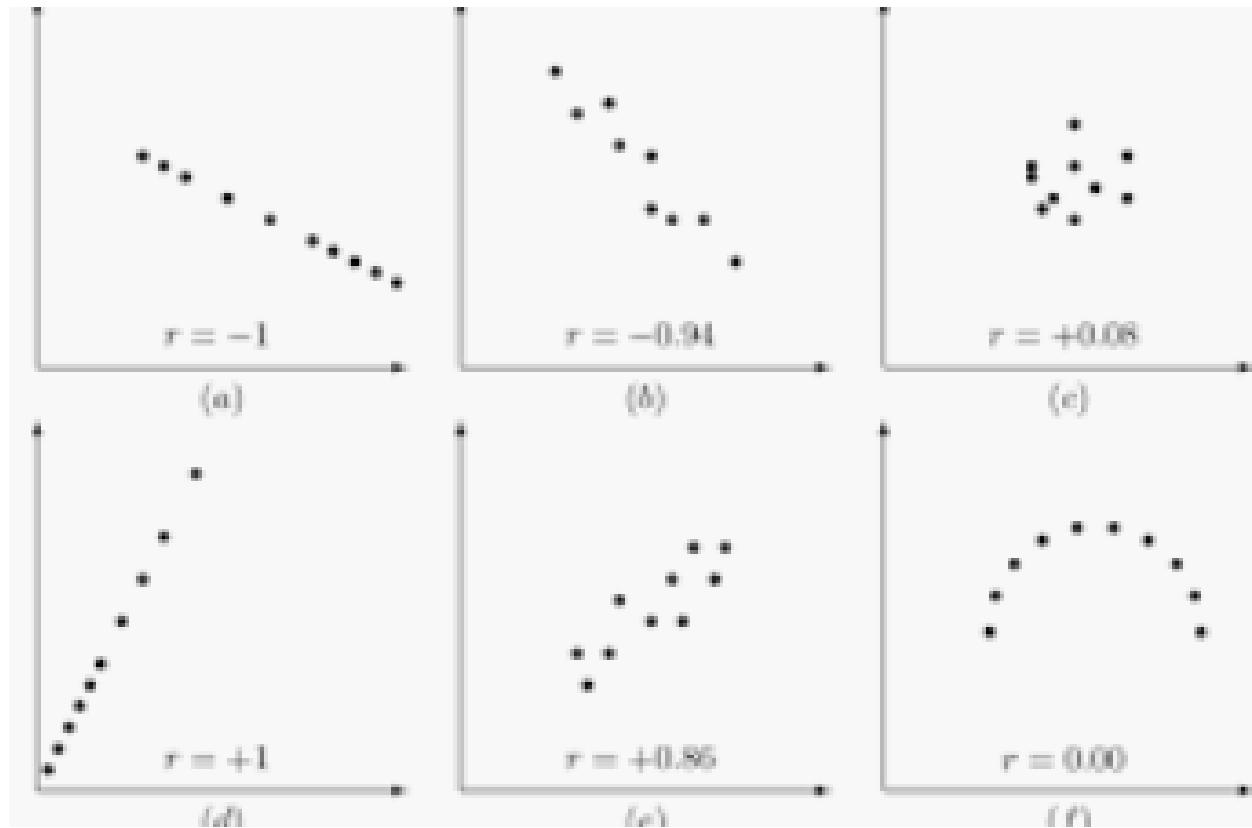
12.5 Correlation

The Sample Correlation Coefficient for the n pairs $(x_1, y_1), \dots, (x_n, y_n)$ is

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx}} \sqrt{s_{yy}}}$$

Properties of r

1. The value of r is independent of the units in which x and y are measured.
2. $-1 \leq r \leq 1$
3. $r = 1$ if and only if all (x_i, y_i) pairs lie on a straight line with positive slope and
 $r = -1$ if and only if all (x_i, y_i) pairs lie on a straight line with negative slope.
5. The square of the sample correlation coefficient gives the value of the coefficient of determination.



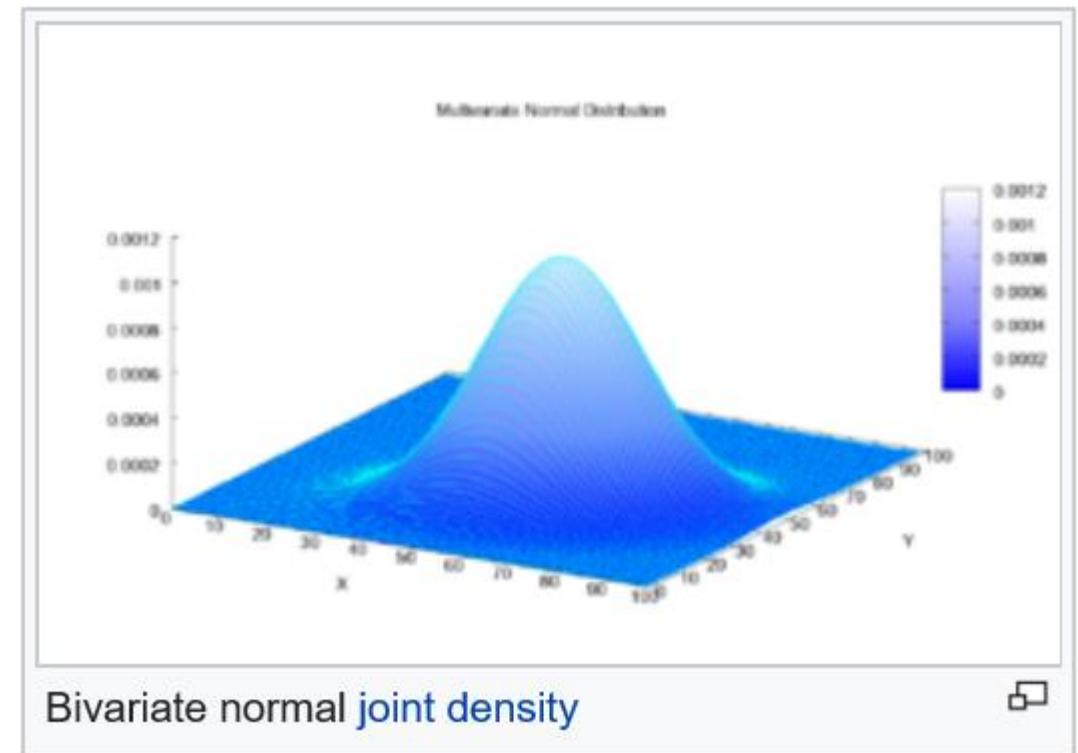
12.5 Correlation

Assumption

The joint probability distribution of (X, Y) is specified by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]/[2(1-\rho^2)]}$$
$$-\infty < X < \infty, -\infty < Y < \infty$$

$f(x, y)$ is called the bivariate normal probability distribution.



Testing for the Absence of Correlation

When : $H_0 : \rho = 0$ is true, the test statistic

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

has a t distribution with $n - 2$ df.

Alternative Hypothesis

$$H_a: \rho > 0$$

$$H_a: \rho < 0$$

$$H_a: \rho \neq 0$$

Rejection Region for Level α Test

$$t \geq t_{\alpha, n-2}$$

$$t \leq -t_{\alpha, n-2}$$

$$\text{either } t \geq t_{\alpha/2, n-2} \text{ or } t \leq -t_{\alpha/2, n-2}$$

A P -value based on $n - 2$ df can be calculated as described previously.

- The relationship between the diameter of a nail(x) and its withdrawal strength(y)

x	2.52	2.87	3.05	3.43	3.68	3.76	3.76	4.50	4.50	5.26
y	54.74	59.01	72.92	50.85	54.99	60.56	69.08	77.03	69.97	90.70

- $H_0 : \rho = 0$ vs $H_a : \rho > 0$

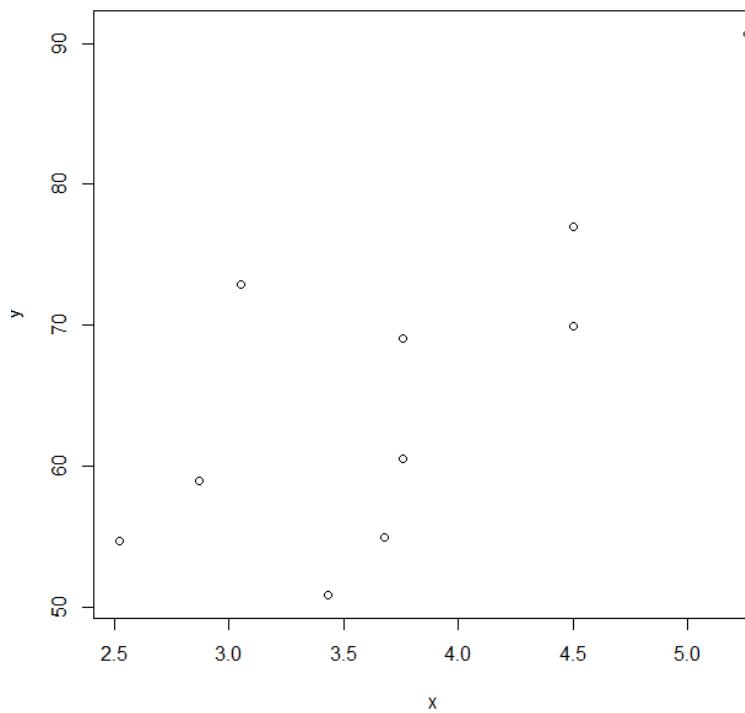
$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7492\sqrt{10-2}}{\sqrt{1-0.7492^2}} = 3.199$$

$$\text{p-value} = P(t_8 > 3.199) = 0.00632$$

Since p-value = 0.00632 < 0.01, it is reasonable to conclude that $\rho > 0$.

```
> 1-pt(3.199, 8) # 0.00632
```

```
> x <- c(2.52, 2.87, 3.05, 3.43, 3.68, 3.76, 3.76, 4.50, 4.50, 5.26)
> y <- c(54.74, 59.01, 72.92, 50.85, 54.99, 60.56, 69.08, 77.03, 69.97, 90.70)
> plot(x, y)
> cor(x, y)
[1] 0.7491725
```



Example

Steps in the analysis

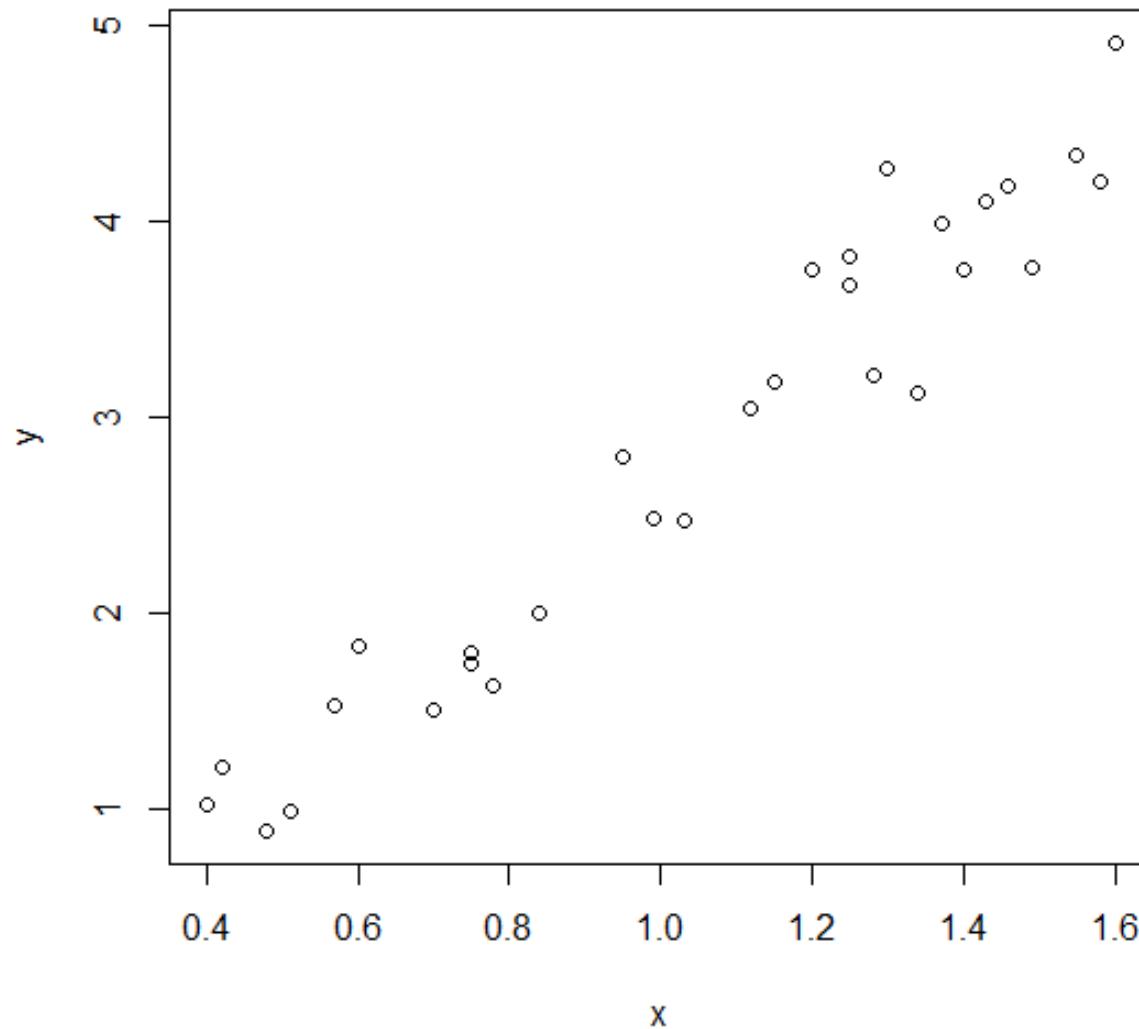
- ① Plot the data
- ② If the linear model is plausible, fit it.
- ③ Check the residual plots and fit a modified model, if necessary.
- ④ Evaluate the significance of the overall model via individual t terms, via t-test s.
- ⑤ Provide predictions and prediction intervals or confidence intervals on the mean response, if required.

Palpebral fissure

Example 12.1 provides measurements of the ocular surface area (OCA) and the palpebral fissure width(the horizontal width of the eye opening) of 30 objects.

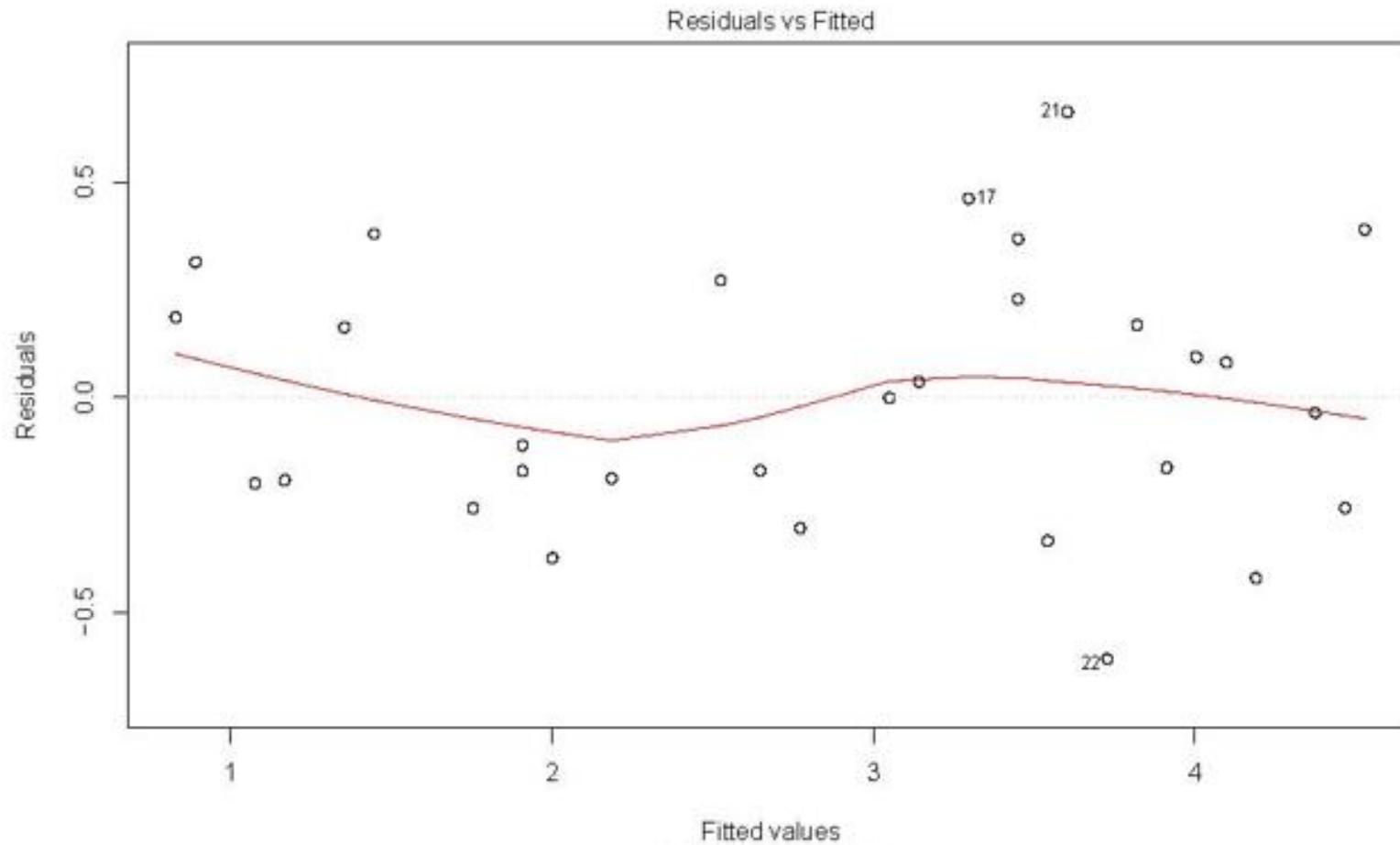
i	1	2	3	4	5	6	7	8	9	10	
	11	12	13	14	15						
x_i	0.40 1.03	0.42 1.12	0.48 0.51	0.51 0.57	0.57 0.60	0.60 0.70	0.70 0.75	0.75 0.78	0.78 0.84	0.84 0.95	0.95 0.99
y_i	1.02 2.48	1.21 2.47	0.88 3.05	0.98 1.52	1.52 1.83	1.83 1.50	1.50 1.80	1.80 1.74	1.74 1.63	1.63 2.00	2.00 2.80
i	16	17	18	19	20	21	22	23	24	25	26
	27	28	29	30							
x_i	1.15 1.55	1.20 1.58	1.25 1.60	1.25 1.28	1.28 1.30	1.30 1.34	1.34 1.37	1.37 1.40	1.40 1.43	1.43 1.46	1.46 1.49
y_i	3.18 4.34	3.76 4.21	3.68 4.92	3.82 3.21	3.21 4.27	4.27 3.12	3.12 3.99	3.99 3.75	3.75 4.10	4.10 4.18	4.18 3.77

```
> with(xmp12.01, plot(y~x))
```



Residual Plot 1 for Palpebral

```
> plot(fm1$fitted.values, fm1$residuals)
```



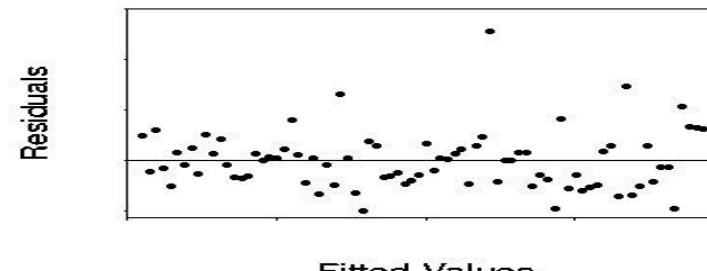
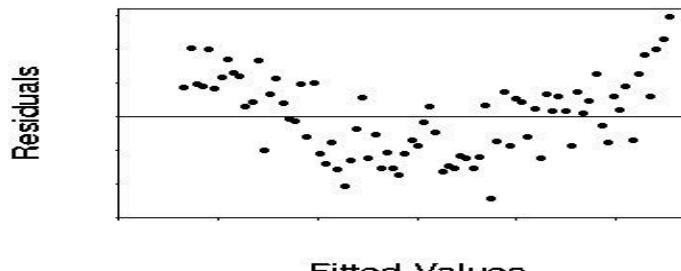
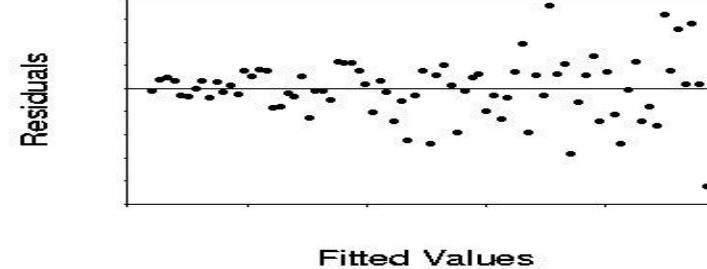
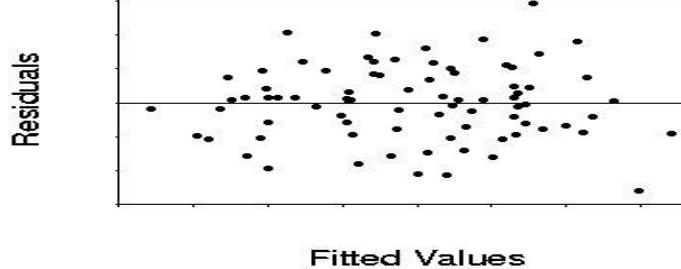
Residual Plots

Upper left: No noticeable pattern.

Upper right: Heteroscedastic.

Lower left: Trend.

Lower Right: Outlier.



Palpebral Fissure Measurement

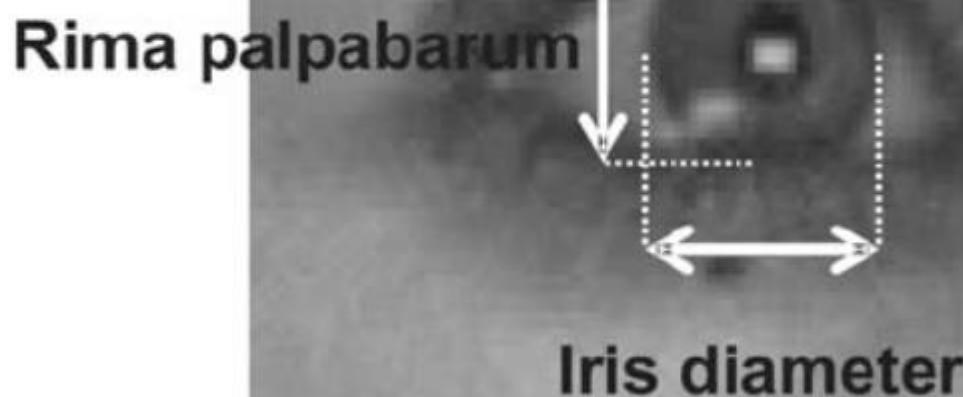
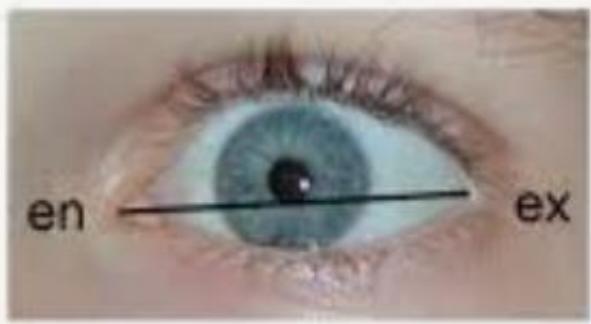


Fig. 2 Calculation of the OSA-proxy by means of the diameter of the iris and the rima palpebrarum

```
> summary(xmp12.01[, 2:3])
```

x	y
Min. :0.400	Min. :0.880
1st Qu.:0.750	1st Qu.:1.755
Median :1.135	Median :3.085
Mean :1.051	Mean :2.840
3rd Qu.:1.363	3rd Qu.:3.808
Max. :1.600	Max. :4.920

```
> fm1 <- lm(y ~ x, data=xmp12.01)
```

```
> summary(fm1)
```

Call:

```
lm(formula = y ~ x, data = xmp12.01)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.60942	-0.19875	-0.01902	0.21727	0.66378

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.3977	0.1680	-2.367	0.0251 *
x	3.0800	0.1506	20.453	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 28 degrees of freedom

Multiple R-squared: 0.9373, Adjusted R-squared: 0.935

F-statistic: 418.3 on 1 and 28 DF, p-value: < 2.2e-16

- Regression equation : Let $\hat{Y} = -0.3977 + 3.08x$
- Residual standard error :

$$s^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{2.6563}{28} = 0.0948$$

$$s = 0.308$$

- Coefficient of determination

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{2.6563}{42.3423} = 0.9373$$

- Standard error of $\hat{\beta}_1$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{0.308}{\sqrt{4.183547}} = 0.1506$$

- 95% CI of $\hat{\beta}_1$

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{\beta}_1} = 3.08 \pm t_{0.025, 28} \cdot 0.1506 = 3.08 \pm t_{0.025, 28} \cdot 0.1506 \rightarrow (2.772, 3.388)$$

- Perform the hypotheses test : $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{3.08}{0.1506} = 20.453$$

$$P\text{-value} = P(t_{30-2} < -20.453) + P(t_{30-2} > 20.453) = 2 \times 1.126 \times 10^{-18} = 2.252 \times 10^{-18}$$

Since $P\text{-value} = 2.252 \times 10^{-18} \leq 0.05$, H_0 can be rejected under the level 0.05 test.

- Compute the 95% CI for $\mu_{Y|1}$, the expected value of Y when $x = 1$

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 \times 1 &\pm t_{0.025, 28} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 \pm t_{0.025, 28} \cdot s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}} \\ &= -0.3977 + (3.08) \times 1 \pm 2.048 \times 0.308 \sqrt{\frac{1}{30} + \frac{(1 - 1.051)^2}{4.184}} \rightarrow (2.566, 2.799)\end{aligned}$$

- Compute the 95% prediction interval for Y when $x = 1$

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 &\pm t_{0.025, 28} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}} \\ &= -0.3977 + (3.08) \times 1 \pm 2.048 \times 0.308 \sqrt{1 + \frac{1}{30} + \frac{(1 - 1.051)^2}{4.184}} \rightarrow (2.041, 3.324)\end{aligned}$$

- Compute the correlation coefficient between x and y.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx}} \sqrt{s_{yy}}} = \frac{12.885}{\sqrt{4.184} \sqrt{42.342}} = 0.9681$$

$$r^2 = 0.9681^2 = 0.9372$$

- Perform the hypotheses test : $H_0 : \rho = 0$ vs $H_a : \rho > 0$

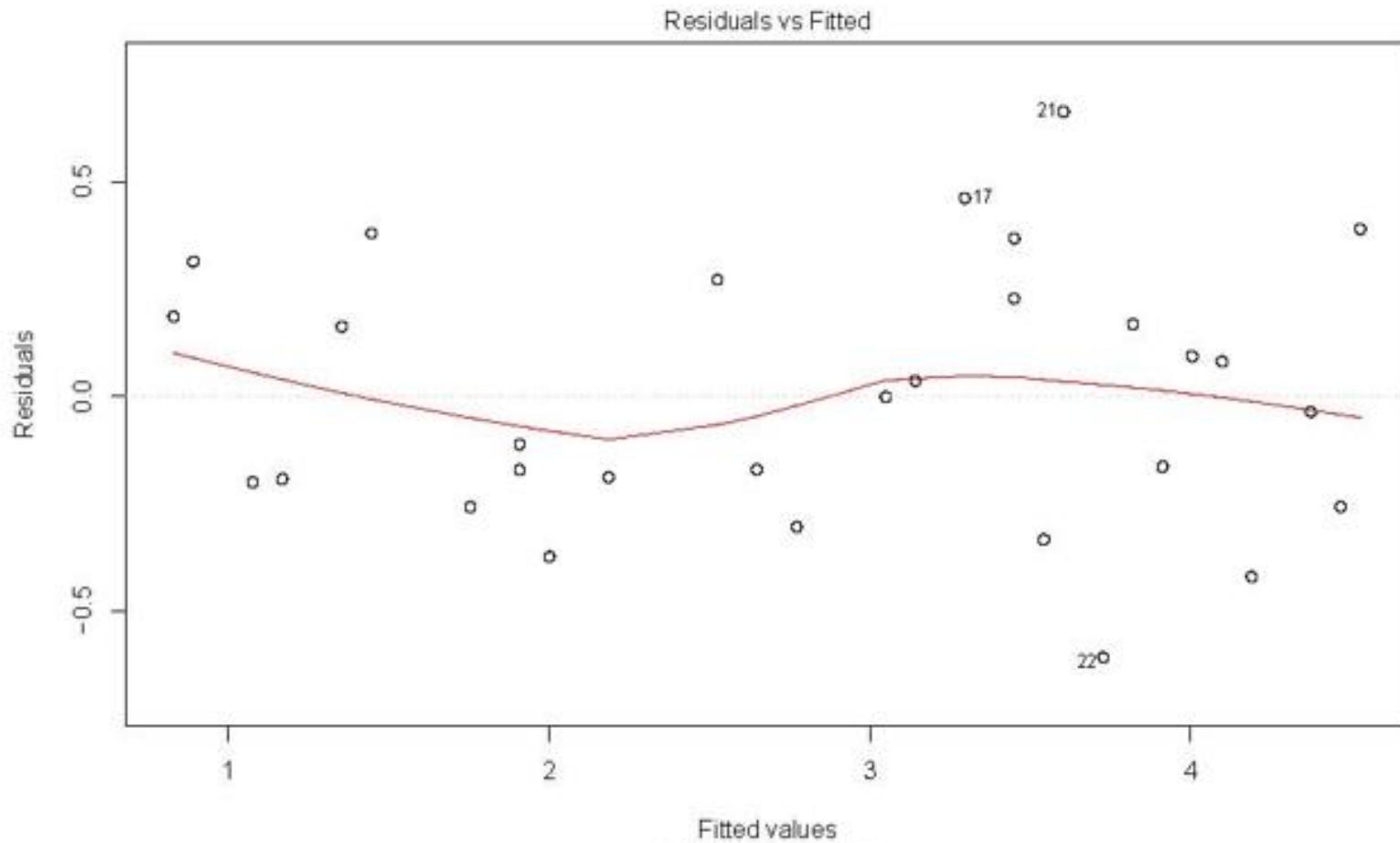
$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9681\sqrt{30-2}}{\sqrt{1-0.9681^2}} = 20.453$$

$$\text{p-value} = P(t_{28} > 20.453) = 0$$

Since p-value = 0 < 0.01, it is reasonable to conclude that $\rho > 0$.

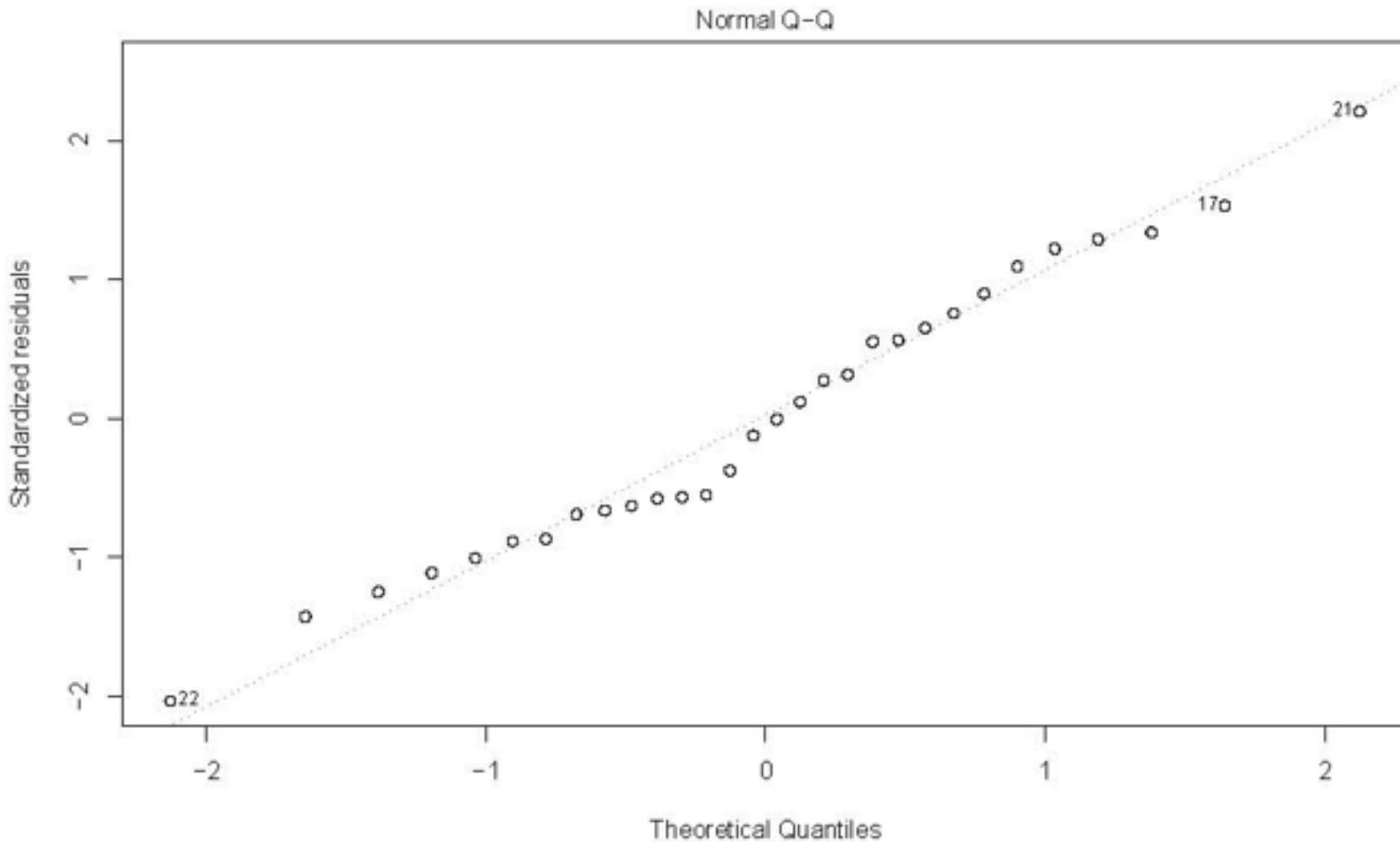
Residual Plot 1 for Palprebal

```
> plot(fm1$fitted.values, fm1$residuals)
```



```
> fm1.stdres = rstandard(fm1)
```

```
> qqnorm(fm1.stdres)
```



- Fitted values

```
> head(fm1$fitted.values, n=10)
```

1	2	3	4	5	6	7
0.8342480	0.8958474	1.0806454	1.1730444	1.3578424	1.4502415	1.7582382
8	9	10				
1.9122365	1.9122365	2.0046356				

- Residuals

```
> head(fm1$residuals, n=10)
```

1	2	3	4	5	6	7
0.1857520	0.3141526	-0.2006454	-0.1930444	0.1621576	0.3797585	-0.2582382
8	9	10				
-0.1122365	-0.1722365	-0.3746356				

Evaluation of predictions

Suppose we wish to evaluate the predictions in the OSA data at 0.6, 1.0, and 1.4.

```
> predict(fm1, list(x=c(0.6, 1, 1.4)))
```

1	2	3
1.450241	2.682228	3.914215

We can evaluate the fitted values and their standard errors as

```
> predict(fm1, list(x=c(0.6, 1, 1.4)), se=TRUE)
```

\$fit

1	2	3
1.450241	2.682228	3.914215

\$se.fit

1	2	3
0.08821348	0.05676328	0.07693567

\$df

[1] 28

\$residual.scale

[1] 0.3080088

Confidence intervals on the mean response

A confidence intervals on the mean response at x^* can be evaluated as

```
> predict(fm1, list(x=c(0.6, 1, 1.4)),interval="conf")
```

	fit	lwr	upr	
1	1.450241	1.269544	1.630939	$1.450241 \pm t_{0.025,28} \times 0.08821348$
2	2.682228	2.565954	2.798503	
3	3.914215	3.756620	4.071811	

Note that the coverage probability of these intervals is for pointwise

A prediction interval (also pointwise coverage) is evaluated as

```
> predict(fm1, list(x=c(0.6, 1, 1.4)),interval="pred")
```

	fit	lwr	upr
1	1.450241	0.7939482	2.106535
2	2.682228	2.0406762	3.323780
3	3.914215	3.2639032	4.564527

Example of Drawing Confidence interval

```
x=c(1,2,3,4,5,6,7,8,9,0)
```

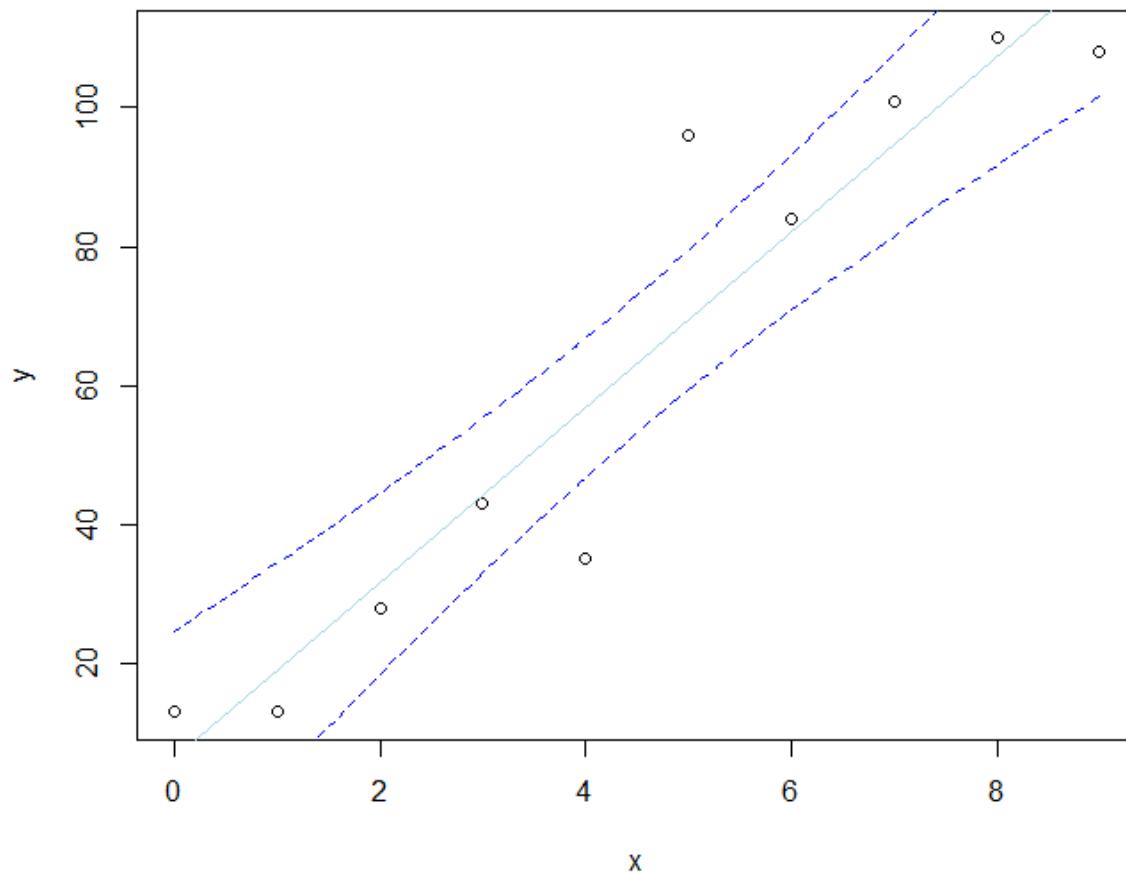
```
y=c(13,28,43,35,96,84,101,110,108,13)
```

```
lm.out <- lm(y ~ x)  
summary(lm.out)
```

```
newx = seq(min(x),max(x),by = 0.05)  
conf_interval <- predict(lm.out, newdata=data.frame(x=newx), interval="confidence",  
level = 0.95)
```

```
plot(x, y, xlab="x", ylab="y", main="Regression")  
abline(lm.out, col="lightblue")  
lines(newx, conf_interval[,2], col="blue", lty=2)  
lines(newx, conf_interval[,3], col="blue", lty=2)
```

Regression



Relationship between income and happiness

<https://www.scribbr.com/statistics/simple-linear-regression/>

You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from \$15k to \$75k and ask them to rank their happiness on a scale from 1 to 10.

```
> income.data <- read.csv("C:\income.data.csv", header=T)
> income.happiness.lm <- lm(happiness ~ income, data = income.data)
> summary(income.happiness.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	0.20427	0.08884	2.299	0.0219 *		
income	0.71383	0.01854	38.505	<2e-16 ***		

signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’	1

Residual standard error: 0.7181 on 496 degrees of freedom

Multiple R-squared: 0.7493, Adjusted R-squared: 0.7488

F-statistic: 1483 on 1 and 496 DF, p-value: < 2.2e-16

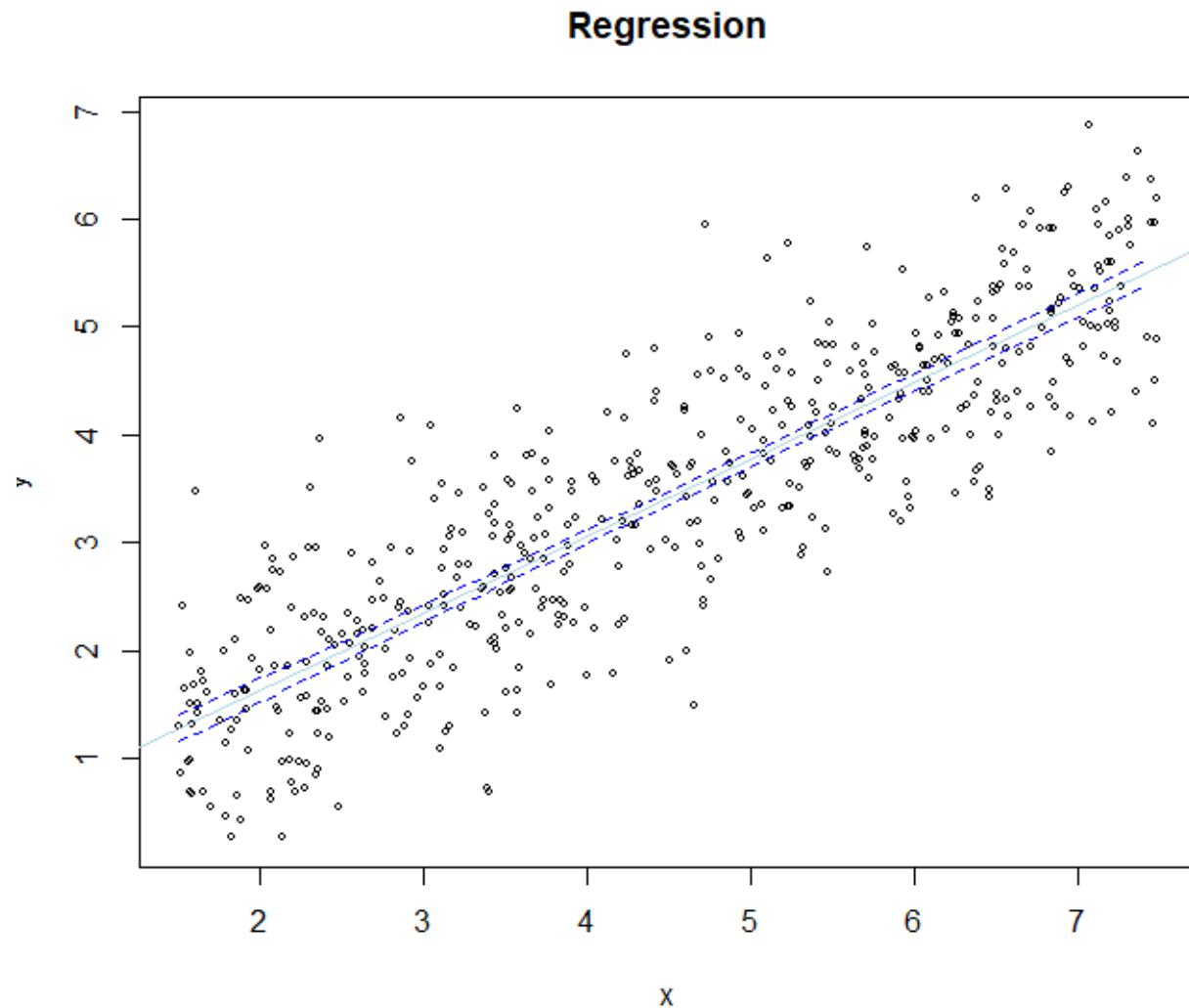
Relationship between income and happiness

```
> income.happiness.lm$coefficients  
(Intercept) income  
0.2042704 0.7138255  
Y= 0.2042704 + 0.7138255 x income
```

The graph of the confidence interval :

```
newx = seq(min(income.data$income),max(income.data$income),by = 0.1)  
newx  
conf_interval <- predict(income.happiness.lm, newdata=data.frame(income=newx),  
                           interval="confidence", level = 0.95)  
plot(income.data$income, income.data$happiness, xlab="x", ylab="y",  
     main="Regression", cex=0.5)  
abline(income.happiness.lm, col="lightblue")  
lines(newx, conf_interval[,2], col="blue", lty=2)  
lines(newx, conf_interval[,3], col="blue", lty=2)
```

Relationship between income and happiness



Chapter 14 – Goodness of Fit Tests and Categorical Data Analysis

Outline

- ① Goodness of Fit Tests When Category Probabilities are Completely Specified
- ② Goodness of Fit for Composite Hypothesis
- ③ Two-Way Contingency Tables

Testing for Homogeneity

Example 14.13 : A company packages a product in cans of 3 different sizes using 3 production line.

Non-conformance reason : Blemish on can, crack in can, improper pull tab location, Pull tab missing, other

Production Line		Reason for Nonconformity					Sample Size
		Blemish	Crack	Location	Missing	Other	
	1	34	65	17	21	13	150
Production Line	2	23	52	25	19	6	125
	3	32	28	16	14	10	100
	Total	89	145	58	54	29	375

$$H_0: p_{1j} = p_{2j} = p_{3j}, \quad j = 1, 2, \dots, 5$$

H_a : the production lines are not homogeneous with respect to the categories

Testing for Independence

Example 14.14 : A study on the relationship between facility conditions of gasoline stations and aggressiveness in the pricing of gasoline was done.

		Observed Pricing Policy			
		Aggressive	Neutral	Non-aggressive	n_i
Condition	Substandard	24	15	17	56
	Standard	52	73	80	205
	Modern	58	86	36	180
	n_j	134	174	133	441
		Expected Pricing Policy			
		Aggressive	Neutral	Non-aggressive	n_i
Condition	Substandard	17.02	22.10	16.89	56
	Standard	62.29	80.88	61.83	205
	Modern	54.69	71.02	54.29	180
	n_j	134	174	133	

14.3 Two-Way Contingency Tables

We consider problems in which the data table have I rows and J columns.

Table 14.9 A Two-Way Contingency Tables

	1	2	...	j	...	J	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	
2	n_{21}						\vdots
\vdots	\vdots						
i	n_{i1}	...		n_{ij}	...		n_i
\vdots	\vdots						
I	n_{I1}	...				n_{IJ}	

$n_{.j}$

Testing for Homogeneity

The proportion of individuals in category j is the same for each population

n_i : the number of sample taken from the i th population

n_{ij} : the number of individuals in the i th sample who fall into category j

$n_j = \sum_{i=1}^I n_{ij}$: the total number of individuals who fall into category j

p_{ij} : the proportion of individuals in population i who fall into category j

\hat{e}_{ij} = estimated expected count in cell (i, j) = $n_i \cdot \frac{n_j}{n}$

$$= \frac{(i\text{th row total})(j\text{th column total})}{n}$$

Testing for Homogeneity

Null hypothesis : $H_0: p_{1j} = p_{2j} = \dots = p_{Ij}, \quad j = 1, 2, \dots, J$

Alternative hypothesis : $H_a: H_0$ is not true

Test statistic value :

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{estimated expected})^2}{\text{estimated expected}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

Rejection Region : $\chi^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$

The test can safely be applied as long as $\hat{e}_{ij} \geq 5$ for all cells.

Testing for Homogeneity

Example 14.13 : A company packages a product in cans of 3 different sizes using 3 production line.

Non-conformance reason : Blemish on can, crack in can, improper pull tab location,
Pull tab missing, other

	Reason for Nonconformity					Sample Size	
	Blemish	Crack	Location	Missing	Other		
Production Line	1	34	65	17	21	13	150
	2	23	52	25	19	6	125
	3	32	28	16	14	10	100
	Total	89	145	58	54	29	375

$$H_0: p_{1j} = p_{2j} = p_{3j}, \quad j = 1, 2, \dots, 5$$

H_a : the production lines are not homogeneous with respect to the categories

Testing for Homogeneity

$$\hat{e}_{11} = \frac{(\text{first row total})(\text{first column total})}{\text{total of sample sizes}} = \frac{(150)(89)}{375} = 35.60$$

$$\hat{e}_{21} = \frac{(\text{second row total})(\text{first column total})}{\text{total of sample sizes}} = \frac{(125)(89)}{375} = 29.67$$

	Production Line	Reason for Nonconformity					Sample Size
		Blemish	Crack	Location	Mission	Other	
1	1	34(35.60)	65(58.00)	17(23.20)	21(21.60)	13(11.60)	150
	2	23(29.67)	52(48.33)	25(19.33)	19(18.00)	6(9.67)	125
	3	32(23.73)	28(38.67)	16(15.47)	14(14.40)	10(7.73)	100
Total		89	145	58	54	29	375

Testing for Homogeneity

The contribution of the (1, 1) cell to χ^2 is then

$$\frac{(\text{observed} - \text{estimated expected})^2}{\text{estimated expected}} = \frac{(34 - 35.60)^2}{35.60} = 0.072$$

$$\begin{aligned}\chi^2 &= \frac{(34 - 35.60)^2}{35.60} + \frac{(65 - 58)^2}{58} + \frac{(17 - 23.2)^2}{23.2} + \dots + \frac{(10 - 7.73)^2}{7.73} \\ &= 0.072 + 0.845 + \dots + 0.664 = 14.159\end{aligned}$$

All estimated expected counts are at least 5.

The test is based on $(3-1)(5-1)=8$ df.

Since $\chi^2 = 14.159 < \chi^2_{0.05,8} = 15.50$, we do not reject H_0 . The production lines are homogeneous. Or

$$P\text{-value} = P(\chi^2_8 > 14.159) = 0.078$$

The null hypothesis of homogeneity should not be rejected at the usual significance levels of 0.05 or 0.01, but it would be rejected for the higher α of 0.10.

```
> 1-pchisq(14.159, 8)
```

```
[1] 0.07771412
```

```
b <- matrix(c(24, 15, 17, 52, 73, 80, 58, 86, 36), nrow=3, byrow=T)  
b  
> chisq.test(b)
```

Pearson's Chi-squared test

```
data: b  
X-squared = 22.476, df = 4, p-value = 0.0001611
```

Testing for Independence

p_{ij} = the proportion of individuals in the population who belong in category i of factor 1
and category j of factor 2

= $P(\text{a randomly selected individual falls in category } i \text{ of factor 1 and category } j \text{ of factor 2})$

$p_{i\cdot} = \sum_{j=1}^J p_{ij} = P(\text{a randomly selected individual falls in category } i \text{ of factor 1})$

$p_{\cdot j} = \sum_{i=1}^I p_{ij} = P(\text{a randomly selected individual falls in category } j \text{ of factor 2})$

$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} = \text{sample proportion for category } i \text{ of factor 1}$

$\hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n} = \text{sample proportion for category } j \text{ of factor 2}$

$\hat{e}_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j} = n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} = \frac{(\text{ith row total})(\text{jth column total})}{n}$

14.3 Two-Way Contingency Tables

We consider problems in which the data table have I rows and J columns.

Table 14.9 A Two-Way Contingency Tables

	1	2	...	j	...	J	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	
2	n_{21}						\vdots
\vdots	\vdots						
i	n_{i1}	...		n_{ij}	...		$n_{i\cdot}$
\vdots	\vdots						
I	n_{I1}	...				n_{IJ}	

$n_{\cdot j}$

Testing for Independence

Null hypothesis : $H_0: p_{ij} = p_{i \cdot} \cdot p_{\cdot j} \quad i = 1, 2, \dots, I \quad j = 1, 2, \dots, J$

Alternative hypothesis : $H_a: H_0$ is not true

Test statistic value :

$$\chi^2 = \sum_{all \ cells} \frac{(observed - estimated \ expected)^2}{estimated \ expected} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

Rejection Region : $\chi^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$

The test can safely be applied as long as $\hat{e}_{ij} \geq 5$ for all cells.

Independence : $P(A \cap B) = P(A)P(B|A) = P(A)P(B)$

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y|X = x) = P(X = x)P(Y = y)$$

Testing for Independence

Example 14.14 : A study on the relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline.

		Observed Pricing Policy			n_i
Condition	Substandard	Aggressive	Neutral	Non-aggressive	
		24	15	17	56
		52	73	80	205
	Modern	58	86	36	180
n_j		134	174	133	441
Expected Pricing Policy					
Condition	Substandard	Aggressive	Neutral	Non-aggressive	n_i
		17.02	22.10	16.89	56
		62.29	80.88	61.83	205
	Modern	54.69	71.02	54.29	180
n_j		134	174	133	441

Testing for Independence

$$\chi^2 = \frac{(24-17.02)^2}{17.02} + \frac{(15-22.10)^2}{22.10} + \dots + \frac{(36-54.29)^2}{54.29} = 22.47$$

All estimated expected counts are at least 5.

The test is based on $(3-1)(3-1)=4$ df.

$$\chi^2_{0.01,4} = 13.277$$

Since $22.47 > 13.277$, the hypothesis of independence is rejected.

$$P\text{-value} = P(\chi_4^2 > 22.47) = 0.00016$$

Since $P\text{-value} = 0.00016 < 0.05$, the hypothesis of independence is rejected at $\alpha = 0.05$

```
> 1-pchisq(22.47, 4)
```

```
[1] 0.00016
```

Chapter 14 – Goodness of Fit Tests and Categorical Data Analysis

Outline

- ① Goodness of Fit Tests When Category Probabilities are Completely Specified
- ② Goodness of Fit for Composite Hypothesis
- ③ Two-Way Contingency Tables

14.1 Goodness of Fit Tests When Category Probabilities are Completely Specified

Null hypothesis : $H_0: p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$

Alternative hypothesis : $H_a:$ at least one p_i does not equal p_{i0}

Test statistic value :

$$\chi^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected} = \sum_{i=1}^k \frac{(n_i - np_{i0})^2}{np_{i0}}$$

Rejection Region : $\chi^2 \geq \chi^2_{\alpha, k-1}$

Goodness of Fit Test (Fitting distributions to the data)

- http://www.cimt.org.uk/projects/mepres/alevel/fstats_ch5.pdf
- Binomial distribution
- Poisson distribution
- Exponential distribution
- Normal distribution : p.107

Binomial Distribution

- Four dice are rolled 200 times. A record is made of the number of dice whose score on the uppermost face is even. The results are as follows.

→ 4

Number of even scores (x_i)	0	1	2	3	4	(2, 3, 5, 6) → 2	(2, 4, 6, 6)
Frequency (f_i)	10	41	70	57	22		

Does the number of dice whose uppermost face is even follow binomial distribution?

(Solution)

$$P(\text{even score on a dice}) = \frac{3}{6} = 0.5$$

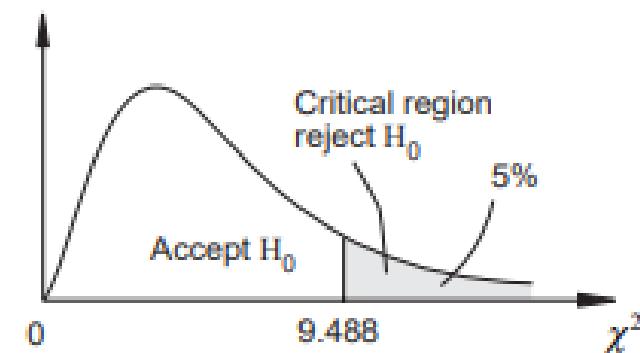
$H_0: X \sim \text{Binom}(4, 0.5)$ vs $H_a: X$ does not follow binomial distribution

$$P(X = 1) = \binom{4}{1} 0.5^1 (1 - 0.5)^3 = 0.25$$

Binomial Distribution

- Use significance level : $\alpha = 0.05$, degrees of freedom $v = 5 - 1 = 4$
(five classes , 1 constraint $\sum E_i = \sum O_i$)

x_i	$O_i = f_i$	$P(X=x_i)$	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	10	0.0625	12.5	-2.5	6.25	0.500
1	41	0.2500	50.0	-9.0	81.00	1.620
2	70	0.3750	75.0	-5.0	25.00	0.333
3	57	0.2500	50.0	7.0	49.00	0.980
4	22	0.0625	12.5	9.5	90.25	7.220
			<u>1.0000</u>		<u>10.653</u>	



Since $\chi^2 = 10.653 > \chi^2_{0.05, 4} = 9.488$, H_0 can be rejected. The number of even scores is not distributed as $\text{Binom}(4, 0.5)$

> qchisq(0.95, 4)

[1] 9.488

Poisson Distribution

- The number of computer malfunctions per day is recorded for 260 days .

Number of Malfunctions (X_i)	0	1	2	3	4	5
Number of days (f_i)	77	90	55	30	5	3
$P(X_i = x)$	0.296	0.346	0.212	0.115	0.019	0.012

Does the number of computer malfunctions per day follow Poisson distribution?
(Solution)

$$\bar{x} = \sum_{x=0}^5 xP(X_i = x) = 1.25 \quad s^2 = \sum_{x=0}^5 x^2 P(X_i = x) - \bar{x}^2 = 1.26$$

$$P(X = x) = \frac{e^{-1.25} 1.25^x}{x!}$$

Poisson Distribution

x_i	$P(X = x_i)$	E_i
0	0.2865	74.5
1	0.3581	93.1
2	0.2238	58.2
3	0.0933	24.2
4	0.0291	7.6
5	0.0073	1.9
$\geq 6^*$	0.0019	0.5
	<u>1.0000</u>	<u>260.0</u>

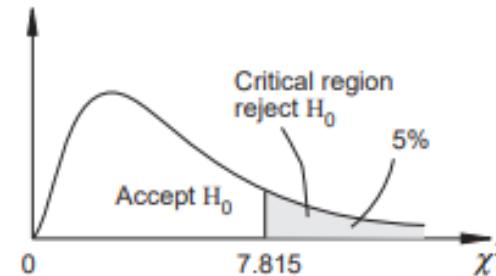
$$260 \times P(X = 1) = 200 \times 0.3581 = 93.1$$

$\chi^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$ is approximated by χ^2 distribution providing none of the expected frequencies are less than 5.
When expected frequencies fall below 5, then groups or classes must be combined.

Poisson Distribution

- Use significance level : $\alpha = 0.05$, degrees of freedom $v = 5 - 2 = 3$
(five classes , 2 constraint $\sum E_i = \sum O_i, \sum E_i x_i = \sum O_i x_i$ from estimation of λ)

x_i	$O_i = f_i$	$P(X=x_i)$	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	77	0.2865	74.5	2.5	6.25	0.084
1	90	0.3581	93.1	-3.1	9.61	0.103
2	55	0.2238	58.2	-3.2	10.24	0.176
3	30	0.0933	24.2	5.8	33.64	1.390
≥ 4	8	0.0383	10.0	-2.0	4.00	0.400
		<u>1.0000</u>				<u>2.153</u>



Since $\chi^2 = 2.153 < \chi^2_{0.05,3} = 7.815$, H_0 cannot be rejected. The number of computer malfunctions per day can be considered to have Poisson distribution.

```
> qchisq(0.95, 3)
```

```
[1] 7.815
```

Normal Distribution

- An analysis of the fat content, $X\%$, of a random sample of 175 hamburgers of a particular grade resulted in the following data

Does the fat content of this grade of hamburger follow normal distribution?

(Solution)

Using class mid-points of 27, 29, ..., 39, we have

$$\bar{x} = \sum_{i=1}^7 x_i P(X = x_i) = 33$$

$$\hat{\sigma}^2 = \sum_{i=1}^7 x_i^2 P(X = x_i) - \bar{x}^2 = 8.411$$

$$\hat{\sigma} = 2.91$$

Use the standardization $z = \frac{x-\mu}{\sigma} = \frac{x-33}{2.91}$

Fat content	Number of hamburgers (f)
$26 \leq x < 28$	7
$28 \leq x < 30$	22
$30 \leq x < 32$	36
$32 \leq x < 34$	45
$34 \leq x < 36$	33
$36 \leq x < 38$	28
$38 \leq x < 40$	4

Normal Distribution

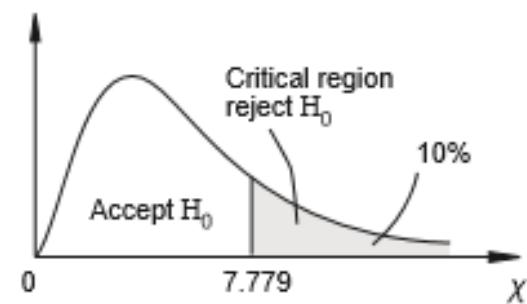
		Class probability	E_i
$P(X < \infty) = P(Z < \infty)$	= 1.000	$P(-\infty < X < 26)$	= 0.008 1.4*
$P(X < 40) = P(Z < 2.405)$	= 0.992	$P(26 \leq X < 28)$	= 0.035 6.1*
$P(X < 38) = P(Z < 1.718)$	= 0.957	$P(28 \leq X < 30)$	= 0.108 18.9
$P(X < 36) = P(Z < 1.031)$	= 0.849	$P(30 \leq X < 32)$	= 0.214 37.5
$P(X < 34) = P(Z < 0.344)$	= 0.635	$P(32 \leq X < 34)$	= 0.270 47.2
$P(X < 32) = P(Z < -0.344)$	= 0.365	$P(34 \leq X < 36)$	= 0.214 37.5
$P(X < 30) = P(Z < -1.031)$	= 0.151	$P(36 \leq X < 38)$	= 0.108 18.9
$P(X < 28) = P(Z < -1.718)$	= 0.043	$P(38 \leq X < 40)$	= 0.035 6.1*
$P(X < 26) = P(Z < -2.405)$	= 0.008	$P(40 \leq X < \infty)$	= 0.008 1.4*
			<hr/> <hr/>
		1.000	175.0

* Combine classes so that all $E_i \geq 5$

Normal Distribution

- Use significance level : $\alpha = 0.10$, degrees of freedom $v = 7 - 3 = 4$
(7 classes , 3 constraint $\sum E_i = \sum O_i$, $\sum E_i x_i = \sum O_i x_i$ from estimation of μ ,
 $\sum E_i x_i^2 = \sum O_i x_i^2$ from estimation of σ^2)

Class	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
$-\infty < x < 28$	7	7.5	-0.5	0.25	0.033
$28 \leq x < 30$	22	18.9	3.1	9.61	0.508
$30 \leq x < 32$	36	37.5	-1.5	2.25	0.060
$32 \leq x < 34$	45	47.2	-2.2	4.84	0.103
$34 \leq x < 36$	33	37.5	-4.5	20.25	0.540
$36 \leq x < 38$	28	18.9	9.1	82.81	4.381
$38 \leq x < \infty$	4	7.5	-3.5	12.25	1.633
					<u>7.258</u>



Since $\chi^2 = 7.258 < \chi^2_{0.10,4} = 7.779$, H_0 cannot be rejected. The fat content of a hamburger can be considered to have normal distribution.

```
> qchisq(0.9, 4)  
[1] 7.779
```