STA 260: Statistics and Probability II

Al Nosedal.
University of Toronto.

Winter 2017

"If you can't explain it simply, you don't understand it well enough"

Albert Einstein.

Toy Problem

- We have a population with a total of five individuals: A, B, C, D, and E .
- We are interested in one variable for this population, X.
- The values of X for this population are: $\{$ 80, 75, 85, 70, 90 $\}$.
- Population average is $\mu = 80$. This is an example of a population parameter.

List of all possible samples

{80,75}	{80,85}	{80,70}
{80,90}	{75,85}	{75,70}
{75,90}	{85,70}	{85,90}
{70,90}		

List of all possible $\bar{X}$s

$$
\begin{array}{lll}
\bar{x}_1 = 77.5 & \bar{x}_2 = 82.5 & \bar{x}_3 = 75 \\
\bar{x}_4 = 85 & \bar{x}_5 = 80 & \bar{x}_6 = 72.5 \\
\bar{x}_7 = 82.5 & \bar{x}_8 = 77.5 & \bar{x}_9 = 87.5 \\
\bar{x}_{10} = 80 & &
\end{array}
$$

Probability distribution for $\bar{X}$

$P(\bar{x} = 72.5) = 1/10$ $\quad P(\bar{x} = 75) = 1/10$ $\quad P(\bar{x} = 77.5) = 2/10$
$P(\bar{x} = 80) = 2/10$ $\quad P(\bar{x} = 82.5) = 2/10$ $\quad P(\bar{x} = 85) = 1/10$
$P(\bar{x} = 87.5) = 1/10$

Expected Value of $\bar{X}$

$$E(\bar{X}) = (72.5)(1/10) + ... + (87.5)(1/10) = 80$$
$$E(\bar{X}^2) = (72.5)^2(1/10) + ... + (87.5)^2(1/10) = 6418.75$$
$$V(\bar{X}) = E(\bar{X}^2) - [E(\bar{X})]^2 = 6418.75 - 6400 = 18.75$$
$$MSE(\bar{X}) = E[(\bar{X} - \mu)^2] = E[\bar{X}^2] - 160E[\bar{X}] + 6400$$
$$MSE(\bar{X}) = 6418.75 - 12800 + 6400 = 18.75$$

Definition 8.1

An **estimator** is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.

Definition 8.2

Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an **unbiased estimator** if $E(\hat{\theta}) = \theta$.

Definition 8.3

The **bias** of a point estimator $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Definition 8.4

The **mean square error** of a point estimator $\hat{\theta}$ is

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Exercise 8.1

Show that

$$MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

Proof

$$\hat{\theta} - \theta = [\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta] = [\hat{\theta} - E(\hat{\theta})] + B(\hat{\theta})$$

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] = E\{[\hat{\theta} - E(\hat{\theta})] + B(\hat{\theta})\}^2 \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2 + [B(\hat{\theta})]^2 + 2B(\hat{\theta})[\hat{\theta} - E(\hat{\theta})]\} \\ &= V(\hat{\theta}) + E\{[B(\hat{\theta})]^2\} + 2B(\hat{\theta})[E(\hat{\theta}) - E(\hat{\theta})] \end{aligned}$$

$$MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

Exercise 8.3

Suppose that $\hat{\theta}$ is an estimator for a parameter θ and $E(\hat{\theta}) = a\theta + b$ for some nonzero constants a and b.

a. In terms of a, b, and θ, what is $B(\hat{\theta})$?

b. Find a function of $\hat{\theta}$ - say, $\hat{\theta}^*$ - that is an unbiased estimator for θ.

Solution

a. By definition

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta = a\theta + b - \theta = (a-1)\theta + b.$$

b. Let $\hat{\theta}^* = \frac{\hat{\theta}-b}{a}$.

$$E(\hat{\theta}^*) = E\left[\frac{\hat{\theta}-b}{a}\right] = \frac{1}{a}E[\hat{\theta} - b] = \theta$$

Exercise 8.5

Refer to Exercises 8.1 and consider the unbiased estimator $\hat{\theta}^*$ that you proposed in Exercise 8.3.

a. Express $MSE(\hat{\theta}^*)$ as a function of $V(\hat{\theta}^*)$.

b. Give an example of a value of a for which $MSE(\hat{\theta}^*) < MSE(\hat{\theta})$.

Solution

a) Note that $E(\hat{\theta}^*) = \theta$ and $V(\hat{\theta}^*) = V\left[\frac{\hat{\theta}-b}{a}\right] = \frac{V(\hat{\theta})}{a^2}$.

$V(\hat{\theta}^*) = \frac{V(\hat{\theta})}{a^2} = MSE(\hat{\theta}^*)$

Solution

b) $MSE(\hat{\theta}) = V(\hat{\theta}) + B^2(\hat{\theta})$

$MSE(\hat{\theta}) = V(\hat{\theta}) + [(a-1)\theta + b]^2$

$MSE(\hat{\theta}^*) = \frac{V(\hat{\theta})}{a^2}$

$MSE(\hat{\theta}^*) < MSE(\hat{\theta})$

$\frac{V(\hat{\theta})}{a^2} < V(\hat{\theta}) + [(a-1)\theta + b]^2$

This last inequality is satisfied for $a > 1$.

Exercise 8.9

Suppose that $Y_1, Y_2, ..., Y_n$ constitute a random sample from a population with probability density function

$$f(y) = \begin{cases} \left(\frac{1}{\theta+1}\right) e^{-y/(\theta+1)} & y > 0,\ \theta > -1 \\ 0 & elsewhere \end{cases}$$

Suggest a suitable statistic to use as an unbiased estimator for θ.

Solution

We know that $\sum_{i=1}^{n} Y_i$ has a Gamma distribution with $\alpha = n$ and $\beta = \theta + 1$.
$E(\sum Y_i) = \alpha\beta = n(\theta + 1) = n\theta + n$
We propose $\hat{\theta}^* = \frac{\sum Y_i - n}{n} = \bar{Y} - 1$
$E(\hat{\theta}^*) = E\left[\frac{\sum Y_i - n}{n}\right] = \frac{1}{n}\left[E(\sum Y_i) - n\right] = \frac{n\theta}{n} = \theta.$

Exercise 8.13

We have seen that if Y has a Binomial distribution with parameters n and p, then Y/n is an unbiased estimator of p. To estimate the variance of Y, we generally use $n(Y/n)(1 - Y/n)$.
a. Show that the suggested estimator is a biased estimator of $V(Y)$.
b. Modify $n(Y/n)(1 - Y/n)$ slightly to form an unbiased estimator of $V(Y)$.

Solution

a) $E(Y) = np$ and $V(Y) = npq$.

$E(Y^2) = npq + (np)^2 = npq + n^2p^2$

$$
\begin{aligned}
E\{n\left(\tfrac{Y}{n}\left(1-\tfrac{Y}{n}\right)\right)\} &= E\{Y - \tfrac{Y^2}{n}\} \\
&= E(Y) - \tfrac{1}{n}E(Y^2) \\
&= np - \tfrac{1}{n}[npq + n^2p^2] \\
&= np - pq - np^2 \\
&= np - p(1-p) - np^2 \\
&= np(1-p) - p(1-p) \\
&= (n-1)p(1-p)
\end{aligned}
$$

Exercise 6.81

Let $Y_1,\ Y_2, ...,\ Y_n$ be independent, exponentially distributed random variables with mean β.
Show that $Y_{(1)} = min(Y_1, Y_2, ..., Y_n)$ has an exponential distribution, with mean β/n.

Solution

Let $U = min(Y_1, Y_2, ..., Y_n)$.
$F_U(u) = P(U \leq u) = P(min(Y_1, Y_2, ..., Y_n) \leq u) =$
$1 - P(min(Y_1, Y_2, ..., Y_n) > u)$
$= 1 - [P(Y_1 > u)P(Y_2 > u)...P(Y_n > u)] = 1 - [P(Y > u)]^n$
$= 1 - [1 - F_Y(u)]^n$
$f_U(u) = \frac{d}{du} F_U(u) = n[1 - F_Y(u)]^{n-1} f_Y(u)$
$f_U(u) = n[1 - 1 + e^{-u/\beta}]^{n-1} \frac{1}{\beta} e^{-u/\beta}$
$f_U(u) = \frac{n}{\beta} e^{-nu/\beta} = \frac{1}{\beta/n} e^{-u/(\beta/n)}$
Clearly, U has an exponential distribution with mean β/n.

Exercise 8.19

Suppose that $Y_1, Y_2, ..., Y_n$ denote a random sample of size n from a population with an exponential distribution whose density is given by

$$f(y) = \begin{cases} (1/\theta)e^{-y/\theta} & y > 0 \\ 0 & elsewhere \end{cases}$$

If $Y_{(1)} = min(Y_1, Y_2, ..., Y_n)$ denotes the smallest-order statistic, show that $\hat{\theta} = nY_{(1)}$ is an unbiased estimator for θ and find $MSE(\hat{\theta})$.

Solution

We know that $U = min(Y_1, Y_2, ..., Y_n)$ has an exponential distribution with mean $\frac{\theta}{n}$.
$E(\hat{\theta}) = E[nU] = nE[U] = n\left(\frac{\theta}{n}\right) = \theta$.
Therefore, $\hat{\theta}$ is unbiased.
Since $\hat{\theta}$ is an unbiased estimator for θ, we have that $MSE(\hat{\theta}) = Var(\hat{\theta})$.
$Var(\hat{\theta}) = Var[nU] = n^2 Var[U] = n^2\left(\frac{\theta}{n}\right)^2 = \theta^2$
Therefore, $MSE(\hat{\theta}) = \theta^2$.

Exercise 8.15

Let $Y_1, \ Y_2, ..., \ Y_n$ denote a random sample of size n from a population whose density is given by

$$f(y) = \begin{cases} 3\frac{\beta^3}{y^4} & \beta \leq y \leq \infty \\ 0 & elsewhere \end{cases}$$

where $\beta > 0$ is unknown. Consider the estimator $\hat{\beta} = min(Y_1, Y_2, ..., Y_n)$.

a. Derive the bias of the estimator $\hat{\beta}$.

b. Derive $MSE(\hat{\beta})$.

Solution

Let $U = min(Y_1, Y_2, ..., Y_n)$
a. From exercise 6.81, we know that
$f_U(u) = n[1 - F_Y(u)]^{n-1} f_Y(u)$.
$F_Y(u) = \int_{\beta}^{u} 3\beta^3 y^{-4} dy = 3\beta^3 \left(\frac{u^{-3}}{-3} - \frac{\beta^{-3}}{3} \right)$
$F_Y(u) = 1 - \frac{\beta^3}{u^3}$
$f_U(u) = n \left(\frac{\beta^3}{u^3} \right)^{n-1} \frac{3\beta^3}{u^4} = 3n \frac{\beta^{3n}}{u^{3n+1}}, \quad \beta \leq u \leq \infty$

Solution

a) $E(U) = \int_{\beta}^{\infty} 3n \frac{\beta^{3n}}{u^{3n+1}} u du = \int_{\beta}^{\infty} 3n \frac{\beta^{3n}}{u^{3n}} du$
$= \frac{3n\beta}{3n-1} \int_{\beta}^{\infty} (3n-1) \frac{\beta^{3n-1}}{u^{3n}} du$

$E(U) = \frac{3n\beta}{3n-1}$

$B(U) = E(U) - \beta = \frac{3n\beta}{3n-1} - \beta = \frac{1}{3n-1}\beta$

Solution

b) $MSE(U) = Var(U) + B^2(U)$

$$E(U^2) = \int_\beta^\infty 3n \frac{\beta^{3n}}{u^{3n+1}} u^2 du = \int_\beta^\infty 3n \frac{\beta^{3n}}{u^{3n-1}} du$$

$$= \frac{3n\beta^2}{3n-2} \int_\beta^\infty (3n-2) \frac{\beta^{3n-2}}{u^{3n-1}} du$$

$$E(U^2) = \frac{3n\beta^2}{3n-2}$$

Solution

$$V(U) = E(U^2) - [E(U)]^2 = \frac{3n\beta^2}{3n-2} - \left(\frac{3n\beta}{3n-1}\right)^2$$

$$B^2(U) = \left(\frac{\beta}{3n-1}\right)^2$$

$$MSE(U) = \frac{3n\beta^2}{3n-2} - \left(\frac{3n\beta}{3n-1}\right)^2 + \left(\frac{\beta}{3n-1}\right)^2$$

$$MSE(U) = \frac{2\beta^2}{(3n-2)(3n-1)}$$

Exercise 8.36

If $Y_1, Y_2, ..., Y_n$ denote a random sample from an exponential distribution with mean θ, then $E(Y_i) = \theta$ and $V(Y_i) = \theta^2$. Suggest an unbiased estimator for θ and provide an estimate for the standard error of your estimator.

Solution

Recall that $U = \sum_{i=1}^{n} Y_i$ has a Gamma distribution with $\alpha = n$ and $\beta = \theta$ (if you don't remember this, show it using the MGF method).
Hence, $E(U) = E(\sum_{i=1}^{n} Y_i) = (\alpha)(\beta) = n\theta$. Thus, we propose $\hat{\theta} = \frac{U}{n} = \bar{Y}$.
$E(\hat{\theta}) = E\left(\frac{U}{n}\right) = \frac{1}{n}E(U) = \theta$.
Clearly, $\hat{\theta}$ is an unbiased estimator for θ.

Solution (cont.)

$Var(\hat{\theta}) = Var\left(\frac{U}{n}\right) = \frac{1}{n^2} Var(U) = \frac{1}{n^2}\alpha\beta^2 = \frac{1}{n^2} n\theta^2 = \frac{\theta^2}{n}$.
We propose $\hat{\sigma}_{\bar{Y}} = \frac{\bar{Y}}{\sqrt{n}}$.

Another solution

Another estimator for θ could be: $\hat{\theta}_2 = Y_1$ (our first observation). Note that it is an unbiased estimator for θ. $E(\hat{\theta}_2) = E(Y_1) = \theta$ and $Var(\hat{\theta}_2) = Var(Y_1) = \theta^2$. Therefore, $\hat{\sigma}_{Y_1} = Y_1$.

Exercise 8.37

Refer to Exercise 8.36. An engineer observes $n = 10$ independent length-of-life measurements on a type of electronic component. The average of these 10 measurements is 1020 hours. If these lengths of life come from an exponential distribution with mean θ, estimate θ and place a 2-standard-error bound on the error of estimation.

Solution

$\hat{\theta}$ = estimate of θ.
$\bar{Y} = \hat{\theta} = 1020$.
2-standard-error bound on the error of estimation:
$2\frac{\bar{Y}}{\sqrt{n}} = 2\frac{1020}{\sqrt{10}} \approx 645.1$

Confidence Interval

Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are the (random) lower and upper confidence limits, respectively, for a parameter θ. Then, if

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

the probability $(1 - \alpha)$ is the confidence coefficient. The resulting random interval defined by $(\hat{\theta}_L, \hat{\theta}_U)$ is called a two-sided confidence interval.

Pivotal quantity

One very useful method for finding confidence intervals is called the **pivotal method**. This method depends on finding a pivotal quantity that possesses two characteristics:

- It is a function of the sample measurements and the unknown parameter θ, where θ is the *only* unknown quantity.
- Its probability distribution does not depend on the parameter θ.

Example

Suppose that we are to obtain a single observation Y from an exponential distribution with mean θ. Use Y to form a confidence interval for θ with confidence coefficient 0.90.

Solution

Let $U = \frac{Y}{\theta}$. Let us find the probability distribution of U.

$$M_U(t) = E[e^{ut}] = E[e^{\frac{Y}{\theta}t}] = E[e^{Y(\frac{t}{\theta})}] = M_Y\left(\frac{t}{\theta}\right)$$

From our table

$$M_Y(\frac{t}{\theta}) = \left[1 - \theta\left(\frac{t}{\theta}\right)\right]^{-1} = [1 - t]^{-1}$$

Clearly, $U = \frac{Y}{\theta}$ has an exponential distribution with mean 1.

Solution

$P[a \leq U \leq b] = 0.90$
Then we would like to find a and b such that
$P[U < a] = P[U \leq a] = 0.05$ and $P[U \leq b] = 0.95$.
That is equivalent to finding a and b such that
$F(a) = 1 - e^{-a} = 0.05$ and $F(b) = 1 - e^{-b} = 0.95$. Solving for a and b yields:
$a = -ln(0.95) = 0.05129$ and $b = -ln(0.05) = 2.9957$. Therefore
$P(0.0513 \leq U \leq 2.996) = 0.90$
$P(0.0513 \leq \frac{Y}{\theta} \leq 2.996) = 0.90$
$P(\frac{Y}{2.996} \leq \theta \leq \frac{Y}{0.0513}) = 0.90$

Exercise 8.39

Suppose that the random variable Y has a Gamma distribution with parameters $\alpha = 2$ and an unknown β. Let $U = \frac{2Y}{\beta}$.
a. Show that U has a χ^2 distribution with 4 degrees of freedom (df).
b. Using $U = \frac{2Y}{\beta}$ as a pivotal quantity, derive a 90% confidence interval for β.

a) Solution

Let Let $U = \frac{2Y}{\beta}$. Let us find the probability distribution of U.

$$M_U(t) = E[e^{ut}] = E[e^{\frac{2Y}{\beta}t}] = E[e^{Y\left(\frac{2t}{\beta}\right)}] = M_Y\left(\frac{2t}{\beta}\right)$$

From our table

$$M_Y\left(\frac{2t}{\beta}\right) = \left[1 - \beta\left(\frac{2t}{\beta}\right)\right]^{-2} = [1 - 2t]^{-4/2}$$

Clearly, $U = \frac{2Y}{\beta}$ has a χ^2 distribution with 4 degrees of freedom (df).

b) Solution

Using table 6 with 4 degrees of freedom,

$$P(0.710721 \leq \frac{2Y}{\beta} \leq 9.48773) = 0.90.$$

So,

$$P\left(\frac{2Y}{9.48773} \leq \beta \leq \frac{2Y}{0.710721}\right) = 0.90$$

and $\left(\frac{2Y}{9.48773}, \frac{2Y}{0.710721}\right)$ forms a 90% CI for β.

Exercise 8.41

Suppose that Y is Normally distributed with mean 0 and unknown variance σ^2. Find a pivotal quantity for σ^2 and use it to give a 95% confidence interval for σ^2.

Example

Let $\hat{\theta}$ be a statistic that is Normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. Find a confidence interval for θ that possesses a confidence coefficient equal to $(1 - \alpha)$.

Solution

Note that $Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ has a Normal distribution with mean 0 and standard deviation 1. Then

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

$$P(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}) = 1 - \alpha$$

$$P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha$$

Exercise 8.59

When it comes advertising, "tweens" are not ready for the hard-line messages that advertisers often use to reach teenagers. The Geppeto Group study found that 78% of tweens understand and enjoy ads that are silly in nature. Suppose that the study involved $n = 1030$ tweens.

a. Construct a 90% confidence interval for the proportion of tweens who understand and enjoy ads that are silly in nature.

b. Do you think that more that 75% of all tweens enjoy ads that are silly in nature? Why?

Solution

a) We know that $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ has, roughly, a Normal distribution with mean 0 and standard deviation 1 (provided that n is "big"). Therefore, a $1-\alpha$ Confidence interval for p is given by:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

$$\left(0.78 - 1.645\sqrt{\frac{(0.78)(0.22)}{1030}}, 0.78 + 1.645\sqrt{\frac{(0.78)(0.22)}{1030}}\right)$$

$$(0.78 - 0.0212, 0.78 + 0.0212)$$

$$(0.7588, 0.8012)$$

Solution

b) The lower endpoint of the interval is 0.7588, so there is evidence that p, the true proportion, is greater than 75%.

Exercise 8.60

What is the normal body temperature for healthy humans? A random sample of 130 healthy human body temperatures provided by Allen Shoemaker yielded 98.25 degrees and standard deviation 0.73 degrees.
a. Give a 99% confidence interval for the average body temperature of healthy people.
b. Does the confidence interval obtained in part a) contain the value 98.6 degrees, the accepted average temperature cited by physicians and others? What conclusions can you draw?

Solution

a) A confidence interval has the form: estimate $\pm$ margin of error. In this case

$$\bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$98.25 \pm 2.57\left(\frac{0.73}{\sqrt{130}}\right)$$

$$98.25 \pm 2.57(0.0640)$$

$$98.25 \pm 0.1645$$

$$(98.0855, 98.4145)$$

Solution

b) Since 98.6 is not included in our interval, we have evidence to claim that the average temperature for healthy humans is different from 98.6 degrees.

Exercise 8.71

A state wildlife service wants to estimate the mean number of days that each licensed hunter actually hunts during a given season, with a bound on the error of estimation equal to 2 hunting days. If data collected in earlier surveys have shown σ to be approximately equal to 10, how many hunters must be included in the survey?

Solution

We know that the margin of error is given by

$$B = z^* \left(\frac{\sigma}{\sqrt{n}} \right).$$

With $B = 2$, $\sigma = 10$, $z^* = 1.96$, and solving for n

$$n = \frac{(z^* \sigma)^2}{B^2} = 97$$

(don't forget, we always round up).

Exercise 8.73

Refer to Exercise 8.59. How many tweens should have been interviewed in order to estimate the proportion of tweens who understand and enjoy ads that are silly in nature, correct to within 0.02 with probability 0.99? Use the proportion from the previous sample in approximating the standard error of the estimate.

Solution

From the previous sample, the proportion of tweens who understand and enjoy ads that are silly in nature is 0.78. Using this as an estimate of p, we estimate the sample size as

$$2.576\sqrt{\frac{(0.78)(1-0.78)}{n}} = 0.02$$

(solving for n)

$$n = 2847$$

Small-Sample Confidence interval for μ

Parameter : μ.

Confidence interval ($\nu =$ df) :

$$\bar{Y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right), \quad \nu = n - 1.$$

Small-Sample Confidence interval for $\mu_1 - \mu_2$

Parameter : $\mu_1 - \mu_2$.

Confidence interval (ν = df) :

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $\nu = n_1 + n_2 - 2$ and $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
(requires that the samples are independent and the assumption that $\sigma_1^2 = \sigma_2^2$).

Definition 7.2

Let Z be a standard Normal random variable and let W be a χ^2-distributed variable with ν df. Then, if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a t distribution with ν df.

Theorem 7.2

Let $Y_1, Y_2, ..., Y_n$ be defined as in Theorem 7.1. Then $Z_i = \frac{Y_i - \mu}{\sigma}$ are independent, standard Normal random variables, $i = 1, 2, ..., n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

Theorem 7.3

Let $Y_1, Y_2, ..., Y_n$ be a random sample from a Normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

has a χ^2 distribution with $(n-1)$ df. Also, $\bar{Y}$ and S^2 are independent random variables.

Development

Let $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ denote a random sample of size n_1 from a population with a Normal distribution with mean μ_1 and variance σ^2. Also, let $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ denote a random sample of size n_2 from a population with a Normal distribution with mean μ_2 and variance σ^2. Then $\bar{Y}_1 - \bar{Y}_2$ has a Normal distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$. This implies that

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a $N(0, 1)$.

Development

The estimator of σ^2 is obtained by pooling the sample data to obtain the *pooled estimator* S_p^2.

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_i^2 is the sample variance from the ith sample, $i = 1, 2$.

Development

Further,

$$W = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2}{\sigma^2} + \frac{\sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2}{\sigma^2}$$

is the sum of two independent χ^2-distributed random variables with $(n_1 - 1)$and $(n_2 - 1)$ df, respectively. Thus, W has a χ^2 distribution with $\nu = (n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$ df. (See Theorems 7.2 and 7.3). We now use the χ^2-distributed variable W and the independent standard normal quantity Z defined above to form a pivotal quantity.

Development

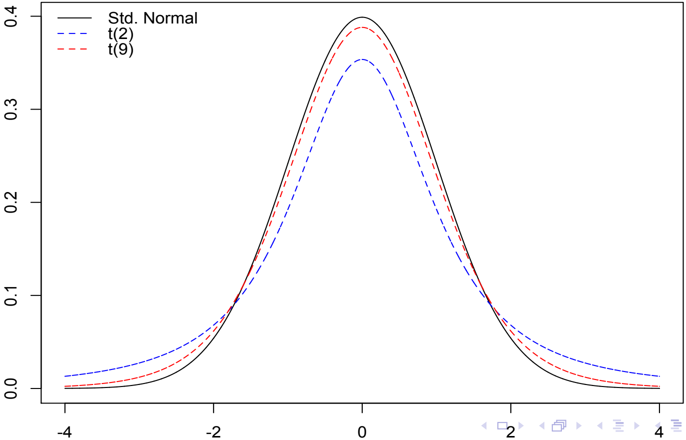
$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

a quantity that by construction has a t distribution with $(n_1 + n_2 - 2)$ df.

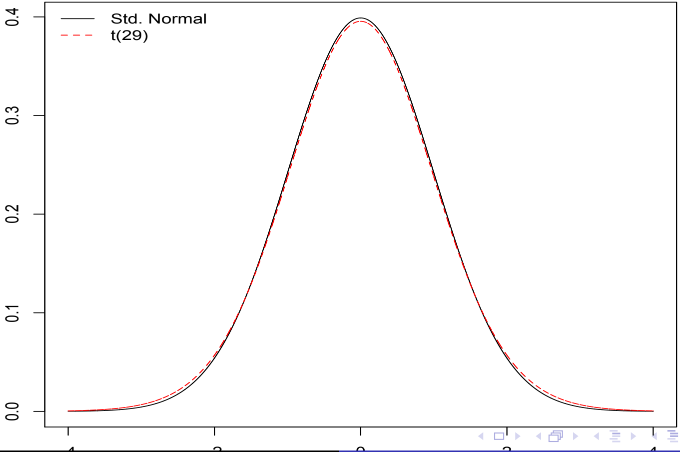
The t distributions

- The density curves of the t distributions are similar in shape to the Standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the t distributions is a bit greater than of the Standard Normal distribution. The t distributions have more probability in the tails and less in the center than does the Standard Normal. This is true because substituting the estimate s for the fixed parameter σ introduces more variation into the statistic.
- As the degrees of freedom increase, the t density curve approaches the $N(0,1)$ curve ever more closely. This happens because s estimates σ more accurately as the sample size increases. So using s in place of σ causes little extra variation when the sample is large.

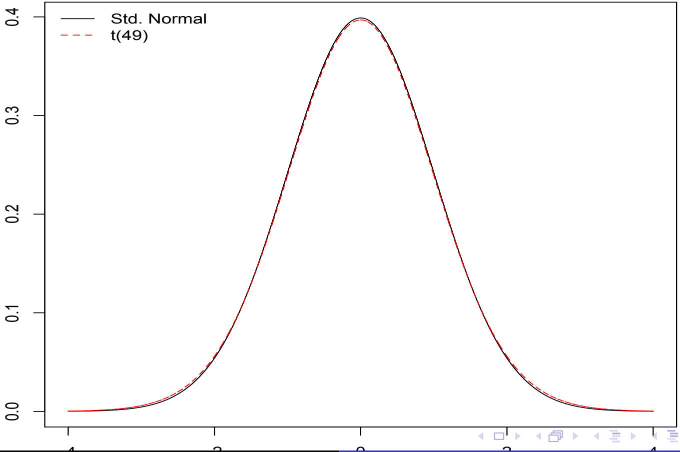
Density curves



Density curves



Density curves



Example: Direct and Broker-Purchased Mutual Funds

Millions of investors buy mutual funds, choosing from thousands of possibilities. Some funds can be purchased directly from banks or other financial institutions whereas others must be purchased through brokers, who charge a fee for this service. This raises the question, Can investors do better by buying mutual funds directly than by purchasing mutual funds through brokers? To help answer this question, a group of researchers randomly sampled the annual returns from mutual funds that can be acquired directly and mutual funds that are bought through brokers and recorded the net annual returns, which are the returns on investment after deducting all relevant fees.

Example: Direct and Broker-Purchased Mutual Funds (cont.)

From the data, the following statistics were calculated:

$n_1 = 50$
$n_2 = 50$
$\bar{x}_1 = 6.63$
$\bar{x}_2 = 3.72$
$s_1^2 = 37.49$
$s_2^2 = 43.34$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The pooled variance estimator is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(49)37.49 + (49)43.34}{50 + 50 - 2} = 40.42$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The number of degrees of freedom of the test statistic is

$$\nu = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The confidence interval estimator of the difference between two means with equal population variance is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

or

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The 95% confidence interval estimate of the difference between the return for directly purchased mutual funds and the mean return for broker-purchased mutual funds is

$$(6.63 - 3.72) \pm 1.984\sqrt{40.42\left(\frac{1}{50} + \frac{1}{50}\right)}.$$

$$2.91 \pm 2.52.$$

The lower and upper limits are 0.39 and 5.43.

Example: Direct and Broker-Purchased Mutual Funds (cont.)

We estimate that the return on directly purchased mutual funds is on average between 0.38 and 5.43 percentage points larger than broker-purchased mutual funds.

Exercise 8.90

Do SAT scores for high school students differ depending on the students' intended field of study? Fifteen students who intended to major in engineering were compared with 15 students who intended to major in language and literature. Given in the accompanying table are the means and standard deviations of the scores on the verbal and mathematics portion of the SAT for the two groups of students:

Exercise 8.90 (cont.)

	Verbal		*Math*	
Engineering	$\bar{y} = 446$	$s = 42$	$\bar{y} = 548$	$s = 42$
Language/Literature	$\bar{y} = 534$	$s = 45$	$\bar{y} = 517$	$s = 52$

Exercise 8.90 (cont.)

a. Construct a 95% confidence interval for the difference in average verbal scores of students majoring in engineering and of those majoring in language/literature.
b. Interpret the results obtained in part a).