Stat 704 Data Analysis I
Probability Review

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

def'n: A **random variable** is defined as a function that maps an outcome from some *random phenomenon* to a real number.

- More formally, a random variable is a map or function from the sample space of an experiment, S, to some subset of the real numbers $R \subset \mathbb{R}$.
- Restated: A random variable assigns a measurement to the result of a random phenomenon.

Example 1: The starting salary (in thousands of dollars) Y for a new tenure-track statistics assistant professor.
Example 2: The number of students N who accept offers of admission to USC's graduate statistics program.

cdf, pdf, pmf

Every random variable has a **cumulative distribution function** (cdf) associated with it:

$$F(y) = P(Y \leq y).$$

Discrete random variables have a probability mass function (pmf)

$$f(y) = P(Y = y) = F(y) - F(y^-) = F(y) - \lim_{x \to y^-} F(x).$$

Continuous random variables have a probability density function (pdf) such that for $a < b$

$$P(a \leq Y \leq b) = \int_a^b f(y)dy.$$

For continuous random variables, $f(y) = F'(y)$.
Question: Are the two examples on the previous slide continuous or discrete?

Example 1

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represent heads.

$$p(x) = (0.5)^x (0.5)^{1-x} \qquad \text{for } x = 0, 1$$

Suppose that we do not know whether or not the coin is fair. Let θ be the probability of a head

$$p(x) = (\theta)^x (1 - \theta)^{1-x} \qquad \text{for } x = 0, 1$$

Example 2

Assume that the time in years from diagnosis until death of a person with a specific kind of cancer follows a density like

$$f(x) = \begin{cases} \frac{e^{-x/5}}{5} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid density?

1. e raised to any power is always positive
2. $\int_0^\infty f(x)\,\mathrm{d}x = \int_0^\infty e^{-x/5}/5\,\mathrm{d}x = -e^{-x/5}\Big|_0^\infty = 1$

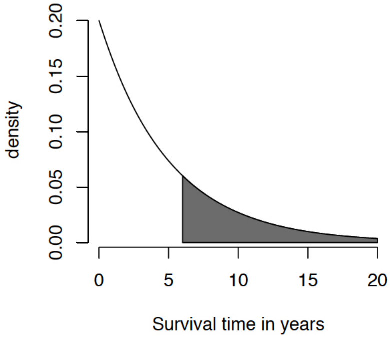
Example 2 (Continued)

What's the probability that a randomly selected person from this distribution survives more than 6 years?

$$P(X \geq 6) = \int_6^\infty \frac{e^{-\frac{t}{5}}}{5}\,\mathrm{d}t = -e^{-t/5}\Big|_6^\infty = e^{-6/5} \approx 0.301$$

Approximiate in R

```
pexp(6, 1/5, lower.tail=FALSE)
```



Quantiles

- The α^{th} **quantiles** of a distribution with distribution function F is the point x_α so that

$$F(x_\alpha) = \alpha$$

- A **percentile** is simple a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

Example 2 (Continued)

- What is the 25^{th} percentile of the exponential distribution considered before? The cumulative density function is

$$F(x) = \int_0^x \frac{e^{-t/5}}{5}\,\mathrm{d}t = -e^{-t/5}\Big|_0^x = 1 - e^{-x/5}$$

- We want to solve for x:

$$\begin{aligned} 0.25 &= F(x) \\ &= 1 - e^{-x/5} \\ x &= -\log(0.75) \times 5 \approx 1.44. \end{aligned}$$

- Therefore, 25% of the patients from this population live less than 1.44 years.
- R can approximate exponential quantile for you

```
qexp(0.25, 1/5)
```

A.3 Expected value

The **expected value**, or **mean** of a random variable is a weighted average according to its probability distribution. It is in general, defined as

$$E\{Y\} = \int_{-\infty}^{\infty} y \, dF(y).$$

For discrete random variables, this is

$$E\{Y\} = \sum_{y:f(y)>0} y \, f(y). \tag{A.12}$$

For continuous random variables this is

$$E\{Y\} = \int_{-\infty}^{\infty} y \, f(y) dy. \tag{A.14}$$

A.3 $E\{\cdot\}$ is linear

Note: If a and c are constants,

$$E\{a + cY\} = a + cE\{Y\}. \tag{A.13}$$

In particular,

$$\begin{aligned} E(a) &= a \\ E\{cY\} &= cE\{Y\} \\ E\{Y + a\} &= E\{Y\} + a \end{aligned}$$

A.3 Variance

The **variance** of a random variable measures the "spread" of its probability distribution. It is the *expected squared deviation about the mean*:

$$\sigma^2\{Y\} = E\{(Y - E\{Y\})^2\} \tag{A.15}$$

Equivalently,

$$\sigma^2\{Y\} = E\{Y^2\} - (E\{Y\})^2 \tag{A.15a}$$

Note: If a and c are constants,

$$\sigma^2\{a + cY\} = c^2\sigma^2\{Y\} \tag{A.16}$$

In particular,

$$\begin{aligned} \sigma^2\{a\} &= 0 \\ \sigma^2\{cY\} &= c^2\sigma^2\{Y\} \\ \sigma^2\{Y + a\} &= \sigma^2\{Y\} \end{aligned}$$

Note: The **standard deviation** of Y is $\sigma\{Y\} = \sqrt{\sigma^2\{Y\}}$.

A.3 Example

Suppose Y is the high temperature in Celsius of a September day in Seattle. Say $E(Y) = 20$ and $\text{var}(Y) = 5$. Let W be the high temperature in Fahrenheit. Then

$$E\{W\} = E\left\{\frac{9}{5}Y + 32\right\} = \frac{9}{5}E\{Y\} + 32 = \frac{9}{5}20 + 32 = 68 \text{ degrees.}$$

$$\sigma^2\{W\} = \sigma^2\left\{\frac{9}{5}Y + 32\right\} = \left(\frac{9}{5}\right)^2 \sigma^2\{Y\} = 3.24(5) = 16.2 \text{ degrees}^2.$$

$$\sigma\{W\} = \sqrt{\sigma^2\{W\}} = \sqrt{16.2} = 4.02 \text{ degrees.}$$

A.3 Covariance

For two random variables Y and Z, the covariance of Y and Z is

$$\sigma\{Y, Z\} = E\{(Y - E\{Y\})(Z - E\{Z\})\}.$$

Note

$$\sigma\{Y, Z\} = E\{YZ\} - E\{Y\}E\{Z\} \tag{A.21}$$

If Y and Z have positive covariance, lower values of Y tend to correspond to lower values of Z (and large values of Y with large values of Z).
Example: Y is work experience in years and Z is salary in €.
If Y and Z have negative covariance, lower values of Y tend to correspond to higher values of Z and vice versa.
Example: Y is the weight of a car in tons and Z is miles per gallon.

A.3 Covariance is linear

If a_1, c_1, a_2, c_2 are constants,

$$\sigma\{a_1 + c_1 Y,\ a_2 + c_2 Z\} = c_1 c_2 \sigma\{Y, Z\} \tag{A.22}$$

Note: by definition $\sigma\{Y, Y\} = \sigma^2\{Y\}$.

The **correlation coefficient** between Y and Z is the covariance scaled to be between -1 and 1:

$$\rho\{Y, Z\} = \frac{\sigma\{Y, Z\}}{\sigma\{Y\}\sigma\{Z\}} \tag{A.25a}$$

If $\rho\{Y, Z\} = 0$ then Y and Z are **uncorrelated**.

A.3 Independent random variables

- Informally, two random variables Y and Z are independent if knowing the value of one random variable does not affect the probability distribution of the other random variable.
- **Note**: If Y and Z are independent, then Y and Z are uncorrelated; i.e., $\rho\{Y, Z\} = 0$.
- However, $\rho\{Y, Z\} = 0$ *does not* imply independence in general.
- If Y and Z have a bivariate normal distribution then $\sigma\{Y, Z\} = 0 \Leftrightarrow Y$, Z independent.
- **Question**: what is the formal definition of independence for (Y, Z)?

Suppose $Y_1, Y_2, \ldots, Y_n$ are random variables and $a_1, a_2, \ldots, a_n$ are constants. Then

$$E\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i E\{Y_i\}. \tag{A.29a}$$

That is,

$$E\{a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n\} = a_1 E\{Y_1\} + a_2 E\{Y_2\} + \cdots + a_n E\{Y_n\}.$$

Also,

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{Y_i, Y_j\} \tag{A.29b}$$

$$
\begin{aligned}
\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} &= E[\sum_{i=1}^{n} a_i Y_i - E(\sum_{i=1}^{n} a_i Y_i)]^2 \\
&= E\{\sum_{i=1}^{n}[a_i Y_i - E(a_i Y_i)]\}^2 = \sum_{i=1}^{n} E\{a_i Y_i - E(a_i Y_i)\}^2 \\
&+ \sum_{i \neq j}^{n} E[a_i Y_i - E(a_i Y_i)][(a_j Y_j) - E(a_j Y_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{Y_i, Y_j\}
\end{aligned}
$$

A.3 Linear combinations of random variables

For two random variables (A.30a & b)

$$\begin{aligned} E\{a_1 Y_1 + a_2 Y_2\} &= a_1 E\{Y_1\} + a_2 E\{Y_2\}, \\ \sigma^2\{a_1 Y_1 + a_2 Y_2\} &= a_1^2 \sigma^2\{Y_1\} + a_2^2 \sigma^2\{Y_2\} + 2 a_1 a_2\, \sigma\{Y_1, Y_2\}. \end{aligned}$$

Note: if $Y_1, \ldots, Y_n$ are all independent (or even just uncorrelated), then

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 \sigma^2\{Y_i\}. \tag{A.31}$$

Also, if $Y_1, \ldots, Y_n$ are all independent, then

$$\sigma\left\{\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i\right\} = \sum_{i=1}^{n} a_i c_i \sigma^2\{Y_i\}. \tag{A.32}$$

A.3 Important example

Suppose $Y_1, \ldots, Y_n$ are independent random variables, each with mean μ and variance σ^2. Define the sample mean as $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then

$$
\begin{aligned}
E\{\bar{Y}\} &= E\left\{\frac{1}{n}Y_1 + \cdots + \frac{1}{n}Y_n\right\} \\
&= \frac{1}{n}E\{Y_1\} + \cdots + \frac{1}{n}E\{Y_n\} \\
&= \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu \\
&= n\left(\frac{1}{n}\mu\right) = \mu.
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2\{\bar{Y}\} &= \sigma^2\left\{\frac{1}{n}Y_1 + \cdots + \frac{1}{n}Y_n\right\} \\
&= \frac{1}{n^2}\sigma^2\{Y_1\} + \cdots + \frac{1}{n^2}\sigma^2\{Y_n\} \\
&= n \times \left(\frac{1}{n^2}\sigma^2\right) = \frac{\sigma^2}{n}.
\end{aligned}
$$

(Casella & Berger pp. 212–214)

A.3 Central Limit Theorem

The **Central Limit Theorem** takes this a step further. When $Y_1, \ldots, Y_n$ are independent and identically distributed (i.e. a *random sample*) from any distribution such that $E\{Y_i\} = \mu$ and $\sigma^2\{Y_i\} = \sigma^2$, and n is reasonably large,

$$\bar{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right),$$

where $\dot{\sim}$ is read as "approximately distributed as".
Note that $E\{\bar{Y}\} = \mu$ and $\sigma^2\{\bar{Y}\} = \frac{\sigma^2}{n}$ as on the previous slide.
The CLT slaps normality onto $\bar{Y}$.
Formally, the CLT states

$$\sqrt{n}(\bar{Y} - \mu) \stackrel{D}{\to} N(0, \sigma^2).$$

(Casella & Berger pp. 236–240)

Normal distribution (Casella & Berger pp. 102–106)

- A random variable Y has a **normal distribution** with mean μ and standard deviation σ, denoted $Y \sim N(\mu, \sigma^2)$, if it has the pdf

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\},$$

  for $-\infty < y < \infty$. Here, $\mu \in \mathbb{R}$ and $\sigma > 0$.
- **Note**: If $Y \sim N(\mu, \sigma^2)$ then $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$ is said to have a **standard normal** distribution.

A.4 Sums of independent normals

Note: If a and c are constants and $Y \sim N(\mu, \sigma^2)$, then

$$a + cY \sim N(a + c\mu, c^2\sigma^2).$$

Note: If $Y_1, \ldots, Y_n$ are independent normal such that $Y_i \sim N(\mu_i, \sigma_i^2)$ and $a_1, \ldots, a_n$ are constants, then

$$\sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + \cdots + a_n Y_n \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Example: Suppose $Y_1, \ldots, Y_n$ are *iid* from $N(\mu, \sigma^2)$. Then

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

(Casella & Berger p. 215)

Exercise

Let

$$Y_{11}, \ldots, Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$$

independent of

$$Y_{21}, \ldots, Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$$

and set $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i,\ i = 1, 2$

1. What is $E\{\bar{Y}_1 - \bar{Y}_2\}$?
2. What is $\sigma^2\{\bar{Y}_1 - \bar{Y}_2\}$?
3. What is the distribution of $\bar{Y}_1 - \bar{Y}_2$?

Exercise

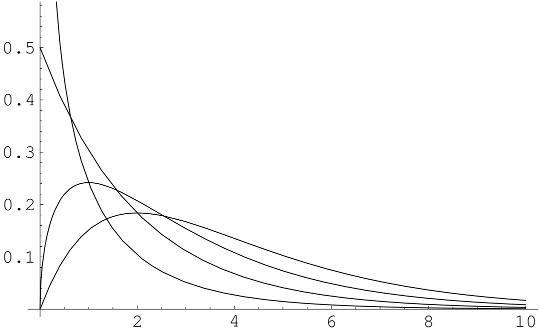
Consider $X \sim N(0,1)$ and $Z \sim N(0,1), X \perp\!\!\!\perp Z$

Let $Y = \rho X + \sqrt{1-\rho^2} Z$

What is

1. $\sigma^2\{Y\}$
2. $\sigma\{X, Y\}$
3. $\rho\{X, Y\}$

def'n: If $Z_1, \ldots, Z_\nu \overset{iid}{\sim} N(0,1)$, then $X = Z_1^2 + \cdots + Z_\nu^2 \sim \chi^2_\nu$, "chi-square with ν degrees of freedom." Note: $E(X) = \nu$ and $\mathrm{var}(X) = 2\nu$. Plot of χ^2_1, χ^2_2, χ^2_3, χ^2_4 PDFs:

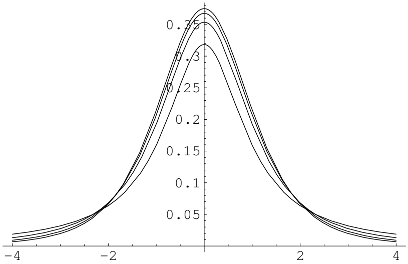


A.4 t distribution

def'n: If $Z \sim N(0,1)$ independent of $X \sim \chi^2_\nu$ then

$$T = \frac{Z}{\sqrt{X/\nu}} \sim t_\nu,$$

"t with ν degrees of freedom."
Note that $E(T) = 0$ for $\nu \geq 2$ and $\mathrm{var}(T) = \frac{\nu}{\nu-2}$ for $\nu \geq 3$.
t_1, t_2, t_3, t_4 PDFs:



A.4 F distribution

def'n: If $X_1 \sim \chi^2_{\nu_1}$ independent of $X_2 \sim \chi^2_{\nu_2}$ then

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

"F with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator."

Note: The square of a t_ν random variable is an $F_{1,\nu}$ random variable. Proof:

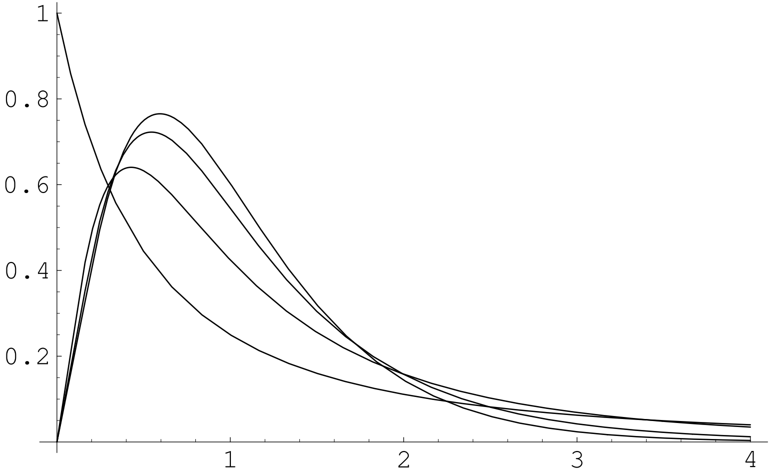
$$t_\nu^2 = \left[\frac{Z}{\sqrt{\chi^2_\nu/\nu}}\right]^2 = \frac{Z^2}{\chi^2_\nu/\nu} = \frac{\chi^2_1/1}{\chi^2_\nu/\nu} = F_{1,\nu}.$$

Note: $E(F) = \nu_2/(\nu_2 - 2)$ for $\nu_2 > 2$. Variance is function of ν_1 and ν_2 and a bit more complicated.

Question: If $F \sim F(\nu_1, \nu_2)$, what is F^{-1} distributed as?

Relate plots to $E(F) = \nu_2/(\nu_2 - 2)$

$F_{2,2}$, $F_{5,5}$, $F_{5,20}$, $F_{5,200}$ PDFs:



A model for a single sample

- Suppose we have a random sample $Y_1, \ldots, Y_n$ of observations from a normal distribution with unknown mean μ and unknown variance σ^2.
- We can model these data as

$$Y_i = \mu + \epsilon_i, \; i = 1, \ldots, n, \text{ where } \epsilon_i \sim N(0, \sigma^2).$$

- Often we wish to obtain inference for the unknown population mean μ, e.g. a confidence interval for μ or hypothesis test $H_0 : \mu = \mu_0$.

A.6 Standardize $\bar{Y}$ to get t random variable

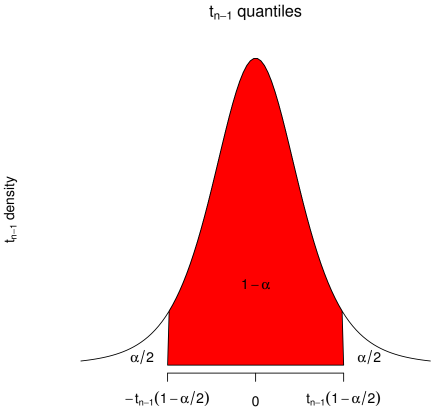
- Let $s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ be the **sample variance** and $s = \sqrt{s^2}$ be the **sample standard deviation.**
- **Fact**: $\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y})^2$ has a χ^2_{n-1} distribution (this can be shown using results from linear models).
- **Fact**: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ has a $N(0,1)$ distribution.
- **Fact**: $\bar{Y}$ is independent of s^2. So then any function of $\bar{Y}$ is independent of any function of s^2.
- Therefore

$$\frac{\left[\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right]}{\sqrt{\frac{\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i-\bar{Y})^2}{n-1}}} = \frac{\bar{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

(Casella & Berger Theorem 5.3.1, p. 218)

A.6 Building a confidence interval

Let $0 < \alpha < 1$, typically $\alpha = 0.05$. Let $t_{n-1}(1 - \alpha/2)$ be such that $P(T \leq t_{n-1}) = 1 - \alpha/2$ for $T \sim t_{n-1}$.



A.6 Confidence interval for μ

Under the model

$$Y_i = \mu + \epsilon_i, \ i = 1, \ldots, n, \text{ where } \epsilon_i \sim N(0, \sigma^2),$$

$$\begin{aligned}
1 - \alpha &= P\left(-t_{n-1}(1-\alpha/2) \le \frac{\bar{Y} - \mu}{s/\sqrt{n}} \le t_{n-1}(1-\alpha/2)\right) \\
&= P\left(-\frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2) \le \bar{Y} - \mu \le \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)\right) \\
&= P\left(\bar{Y} - \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2) \le \mu \le \bar{Y} + \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)\right)
\end{aligned}$$

A.6 Confidence interval for μ

So a $(1-\alpha)100\%$ *random* probability interval for μ is

$$\bar{Y} \pm t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}}$$

where $t_{n-1}(1-\alpha/2)$ is the $(1-\alpha/2)$th quantile of a t_{n-1} random variable: i.e. the value such that $P(T < t_{n-1}(1-\alpha/2)) = 1-\alpha/2$ where $T \sim t_{n-1}$.

This, of course, turns into a “confidence interval” after $\bar{Y} = \bar{y}$ and s^2 are observed, and no longer random.

A.6 Standardizing with $\bar{Y}$ instead of μ

Note: If $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then:

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

and

$$\sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{\sigma} \right)^2 \sim \chi^2_{n-1}.$$

Confidence interval example

Say we collect $n = 30$ summer daily high temperatures and obtain $\bar{y} = 77.667$ and $s = 8.872$. To obtain a 90% CI, we need, where $\alpha = 0.10$,

$$t_{29}(1 - \alpha/2) = t_{29}(0.95) = 1.699,$$

yielding

$$77.667 \pm (1.699)\left(\frac{8.872}{\sqrt{30}}\right) \Rightarrow (74.91, 80.42).$$

Interpretation:
With 90% confidence, the interval between 74.91 and 80.42 degrees covers the true mean high temperature.

Example: Page 645

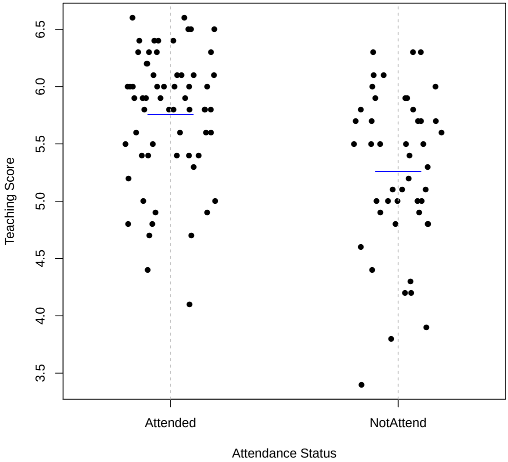
- $n = 63$ faculty voluntarily attended a summer workshop on case teaching methods (out of 110 faculty total).
- At the end of the following academic year, their teaching was evaluated on a 7-point scale (1=really bad to 7=outstanding).
- Calculate the mean and confidence interval for the mean for only the “Attended cases.”

R code

```
##############################
# Example 2, p. 645 (Chapter 15)
###############################

scores<-c(4.8, 6.4, 6.3, 6.0, 5.4,
         5.8, 6.1, 6.3, 5.0, 6.2,
         5.6, 5.0, 6.4, 5.8, 5.5,
         6.1, 6.0, 6.0, 5.4, 5.8,
         6.5, 6.0, 6.1, 4.7, 5.6,
         6.1, 5.8, 4.8, 5.9, 5.4,
         5.3, 6.0, 5.6, 6.3, 5.2,
         6.0, 6.4, 5.8, 4.9, 4.1,
         6.0, 6.4, 5.9, 6.6, 6.0,
         4.4, 5.9, 6.5, 4.9, 5.4,
         5.8, 5.6, 6.2, 6.3, 5.8,
         5.9, 6.5, 5.4, 5.9, 6.1,
         6.6, 4.7, 5.5, 5.0, 5.5,
         5.7, 4.3, 4.9, 3.4, 5.1,
         4.8, 5.0, 5.5, 5.7, 5.0,
         5.2, 4.2, 5.7, 5.9, 5.8,
         4.2, 5.7, 4.8, 4.6, 5.0,
         4.9, 6.3, 5.6, 5.7, 5.1,
         5.8, 3.8, 5.0, 6.1, 4.4,
         3.9, 6.3, 6.3, 4.8, 6.1,
         5.3, 5.1, 5.5, 5.9, 5.5,
         6.0, 5.4, 5.9, 5.5, 6.0)

status<-c(rep("Attended", 63), rep("NotAttend", 47))
dat<-data.frame(scores, status)
str(dat)
```

R code

```
############ make a side-by-side dotplot
keep<-which(dat[,2]=="Attended")
pdf("Teaching.pdf")
stripchart(dat[,1] ~ dat[,2], pch=21, method="jitter", jitter=0.2, vertical=TRUE,
ylab="Teaching Score",  xlab="Attendance Status", ylim=c(min(dat[,1]), max(dat[,1])))
abline(v=1, col="grey", lty=2)
abline(v=2, col="grey", lty=2)
lines(x=c(0.9, 1.1), rep(mean(dat[keep,1]),2), col=4)
lines(x=c(1.9, 2.1), rep(mean(dat[-keep,1]),2), col=4)
dev.off()

########### calculate CI for scores from Attended cases
dat2<-dat[keep,]
str(dat2)
t.test(dat2[,1])
```



Teaching Score
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Attended
NotAttend
Attendance Status