Stat 704 Data Analysis I
Probability Review

Timothy Hanson

Department of Statistics, University of South Carolina

Course information

- Logistics: Tuesday/Thursday 11:40am to 12:55pm in LeConte College 201A.
- Instructor: Tim Hanson, Leconte 219C.
- Office hours: Tuesday/Thursday 10-11:00am and by appointment.
- Required text: *Applied Linear Statistical Models* (5th Edition), by Kutner, Nachtsheim, Neter, and Li.
- Online notes at http://www.stat.sc.edu/∼hansont/stat704/stat704.html
- Grading: homework 50%, two exams 25% each.
- Stat 704 has a co-requisite of Stat 712 (Casella & Berger level mathematical statistics). You need to be taking this, or have taken this already.

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 measures the result of a random phenomenon.

Example 1: The height Y of a randomly selected University of South Carolina statistics graduate student.
Example 2: The number of car accidents Y in a month at the intersection of Assembly and Gervais.

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)

$$f(y) = P(Y = y) = F(y) - F(y-) = F(y) - \lim_{x \to y^-} F(x). \quad \text{(A.11)}$$

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?

A.3 Expected value

The **expected value**, or **mean** of a random variable is, in general, defined as

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

For discrete random variables this is

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

For continuous random variables this is

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

$E(\cdot)$ 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*:

$$\text{var}(Y) = E\{[Y - E(Y)]^2\} \tag{A.15}$$

Equivalently,

$$\text{var}(Y) = E(Y^2) - [E(Y)]^2 \tag{A.15a}$$

Note: If a and c are constants,

$$\text{var}(a + cY) = c^2\text{var}(Y) \tag{A.16}$$

In particular,

$$\begin{aligned} \text{var}(a) &= 0 \\ \text{var}(cY) &= c^2\text{var}(Y) \\ \text{var}(Y + a) &= \text{var}(Y) \end{aligned}$$

Note: The **standard deviation** of Y is $\text{sd}(Y) = \sqrt{\text{var}(Y)}$.

Example

Suppose Y is the high temperature in Celsius of a September day in Seattle. Say $E(Y) = 20$ and $\text{var}(Y) = 10$. Let W be the high temperature in Fahrenheit. Then

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

$$\text{var}(W) = \text{var}\left(\frac{9}{5}Y + 32\right) = \left(\frac{9}{5}\right)^2 \text{var}(Y) = 3.24(10) = 32.4 \text{ degrees}^2.$$

$$\text{sd}(Y) = \sqrt{\text{var}(Y)} = \sqrt{32.4} = 5.7 \text{ degrees.}$$

A.3 Covariance

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

$$\mathrm{cov}(Y, Z) = E\{[Y - E(Y)][Z - E(Z)]\}.$$

Note

$$\mathrm{cov}(Y, Z) = E(YZ) - E(Y)E(Z) \tag{A.21}$$

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: X is work experience in years and Y is salary in Euro.

If Y and Z have negative covariance, lower values of Y tend to correspond to higher values of Z and vice-versa.

Example: X is the weight of a car in tons and Y is miles per gallon.

Covariance is linear

If a_1, c_1, a_2, c_2 are constants,

$$\mathrm{cov}(a_1 + c_1 Y,\ a_2 + c_2 Z) = c_1 c_2 \mathrm{cov}(Y, Z) \tag{A.22}$$

Note: by definition $\mathrm{cov}(Y, Y) = \mathrm{var}(Y)$.

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

$$\mathrm{corr}(Y, Z) = \frac{\mathrm{cov}(Y, Z)}{\sqrt{\mathrm{var}(Y)\mathrm{var}(Z)}} \tag{A.25a}$$

If $\mathrm{corr}(Y, Z) = 0$ then Y and Z are **uncorrelated**.

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, $\mathrm{corr}(Y, Z) = 0$.
- However, $\mathrm{corr}(Y, Z) = 0$ *does not* imply independence in general.
- If Y and Z have a bivariate normal distribution then $\mathrm{cov}(Y, Z) = 0 \Leftrightarrow Y, Z$ independent.
- **Question**: what is the formal definition of independence for (Y, Z)?

Linear combinations of random variables

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,

$$E\left[a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n\right] = a_1 E(Y_1) + a_2 E(Y_2) + \cdots + a_n E(Y_n).$$

Also,

$$\text{var}\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \text{cov}(Y_i, Y_j) \tag{A.29b}$$

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

$$\begin{aligned} E(a_1 Y_1 + a_2 Y_2) &= a_1 E(Y_1) + a_2 E(Y_2), \\ \text{var}(a_1 Y_1 + a_2 Y_2) &= a_1^2 \text{var}(Y_1) + a_2^2 \text{var}(Y_2) + 2 a_1 a_2 \text{cov}(Y_1, Y_2). \end{aligned}$$

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

$$\text{var}\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i^2 \text{var}(Y_i). \qquad \text{(A.31)}$$

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

$$\text{cov}\left(\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} a_i c_i \text{var}(Y_i). \qquad \text{(A.32)}$$

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

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

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

(Casella & Berger pp. 212–214)

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 $\mathrm{var}(Y) = \sigma^2$, and n is reasonably large,

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

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

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

(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.

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

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

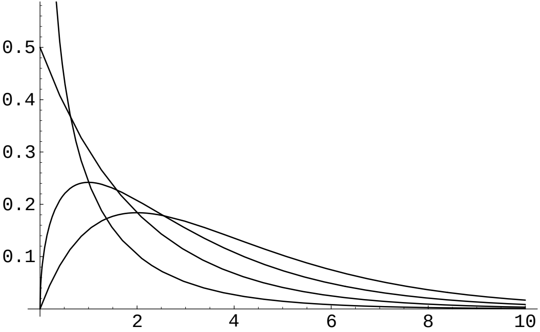
Example: Suppose $Y_1, \ldots, Y_n$ are *iid* from $N(\mu, \sigma^2)$. Then

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

(Casella & Berger p. 215)

A.4 χ^2 distribution

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: $E(X) = \nu$ & $\mathrm{var}(X) = 2\nu$. Plot of χ^2_1, χ^2_2, χ^2_3, χ^2_4 PDFs:

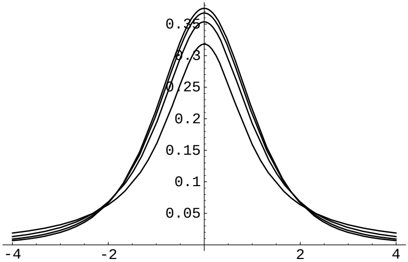


A.4 t distribution

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:

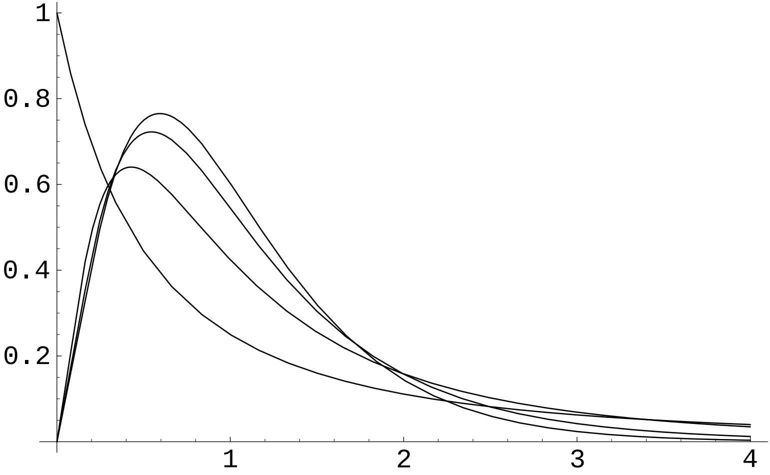
$$t_\nu^2 = \left[\frac{Z}{\sqrt{\chi^2_\nu/\nu}}\right]^2 = \frac{Z^2}{\chi^2_\nu/\nu} = \frac{\chi^2_1/1}{\chi^2_\nu/\nu} = F_{1,\nu}.$$

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 $H_0 : \mu = \mu_0$.

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

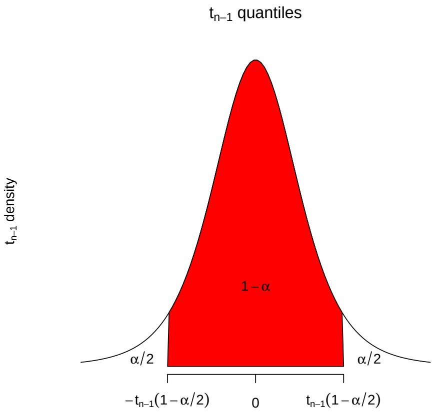
- Let $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ be the **sample variance** and $s = \sqrt{s^2}$ be the **sample standard deviation**.
- **Fact**: $\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ has a χ^2_{n-1} distribution (easy to show using results from linear 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)

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 $T \sim t_{n-1}$.



Confidence interval for μ

Under the model

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

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

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.

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

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

First one is straightforward from properties of normals and definition of χ^2_ν; second one is intuitive but *not* straightforward to show until linear models...

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 $\alpha = 0.10$

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

yielding

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

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

SAS 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).
- `proc ttest` in SAS gets us a confidence interval for the mean.

SAS code

```
*******************************
* Example 2, p. 645 (Chapter 15)
*******************************;
data teaching;
input rating attend$ @@;
if attend='Attended'; * only keep those who attended;
datalines;
 4.8  Attended   6.4  Attended   6.3  Attended   6.0  Attended   5.4  Attended
 5.8  Attended   6.1  Attended   6.3  Attended   5.0  Attended   6.2  Attended
 5.6  Attended   5.0  Attended   6.4  Attended   5.8  Attended   5.5  Attended
 6.1  Attended   6.0  Attended   6.0  Attended   5.4  Attended   5.8  Attended
 6.5  Attended   6.0  Attended   6.1  Attended   4.7  Attended   5.6  Attended
 6.1  Attended   5.8  Attended   4.8  Attended   5.9  Attended   5.4  Attended
 5.3  Attended   6.0  Attended   5.6  Attended   6.3  Attended   5.2  Attended
 6.0  Attended   6.4  Attended   5.8  Attended   4.9  Attended   4.1  Attended
 6.0  Attended   6.4  Attended   5.9  Attended   6.6  Attended   6.0  Attended
 4.4  Attended   5.9  Attended   6.5  Attended   4.9  Attended   5.4  Attended
 5.8  Attended   5.6  Attended   6.2  Attended   6.3  Attended   5.8  Attended
 5.9  Attended   6.5  Attended   5.4  Attended   5.9  Attended   6.1  Attended
 6.6  Attended   4.7  Attended   5.5  Attended   5.0  NotAttend  5.5  NotAttend
 5.7  NotAttend  4.3  NotAttend  4.9  NotAttend  3.4  NotAttend  5.1  NotAttend
 4.8  NotAttend  5.0  NotAttend  5.5  NotAttend  5.7  NotAttend  5.0  NotAttend
 5.2  NotAttend  4.2  NotAttend  5.7  NotAttend  5.9  NotAttend  5.8  NotAttend
 4.2  NotAttend  5.7  NotAttend  4.8  NotAttend  4.6  NotAttend  5.0  NotAttend
 4.9  NotAttend  6.3  NotAttend  5.6  NotAttend  5.7  NotAttend  5.1  NotAttend
 5.8  NotAttend  3.8  NotAttend  5.0  NotAttend  6.1  NotAttend  4.4  NotAttend
 3.9  NotAttend  6.3  NotAttend  6.3  NotAttend  4.8  NotAttend  6.1  NotAttend
 5.3  NotAttend  5.1  NotAttend  5.5  NotAttend  5.9  NotAttend  5.5  NotAttend
 6.0  NotAttend  5.4  NotAttend  5.9  NotAttend  5.5  NotAttend  6.0  NotAttend
;
proc ttest data=teaching;
 var rating;
run;
```