Course information:

- Instructor: Tim Hanson, Leconte 219C, phone 777-3859.
- Office hours: Tuesday/Thursday 11-12, Wednesday 10-12, and by appointment.
- 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 based on David Hitchcock's notes and the text.
- Grading, et cetera: see syllabus.
- 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.

Section A.3 Random Variables

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.

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

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

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 $a < b$

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

For continuous random variables, $f(y) = F'(y)$.

Question: Are the two examples on the previous slide continuous or discrete?

Expected value (Casella & Berger 2.3, 2.3)

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

$$E(Y) = \sum_{y:f(y)>0} y \; f(y).$$

For continuous random variables this is

$$E(Y) = \int_{-\infty}^{\infty} y \; f(y) dy.$$

Note: If a and c are constants,

$$E(a + cY) = a + cE(Y).$$

In particular,

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

Variance

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

$$\mathrm{var}(Y) = E\{[Y - E(Y)]^2\}.$$

Equivalently,

$$\mathrm{var}(Y) = E(Y^2) - [E(Y)]^2.$$

Note: If a and c are constants, $\mathrm{var}(a + cY) = c^2\mathrm{var}(Y)$. In particular,

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

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

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

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

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

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

Covariance (C & B Section 2.5)

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

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

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.

If a_1, c_1, a_2, c_2 are constants,

$$\text{cov}(a_1 + c_1 Y,\ a_2 + c_2 Z) = c_1 c_2 \text{cov}(Y, Z).$$

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

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

$$\text{corr}(Y, Z) = \frac{\text{cov}(Y, Z)}{\sqrt{\text{var}(Y)\text{var}(Z)}}.$$

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

Independent random variables (C & B 4.2)

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, $\text{corr}(Y, Z) = 0$.

However, $\text{corr}(Y, Z) = 0$ *does not* imply independence in general.

If Y and Z have a bivariate normal distribution then $\text{cov}(Y, Z) = 0$ $\Leftrightarrow$ 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

$$E\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i E(Y_i).$$

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,

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

For two random variables

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

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

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

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

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

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

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

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

(C & B p. 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

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

(C & B pp. 236–240)

Section A.4 Gaussian & related distributions

Normal distribution (C & B 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.

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

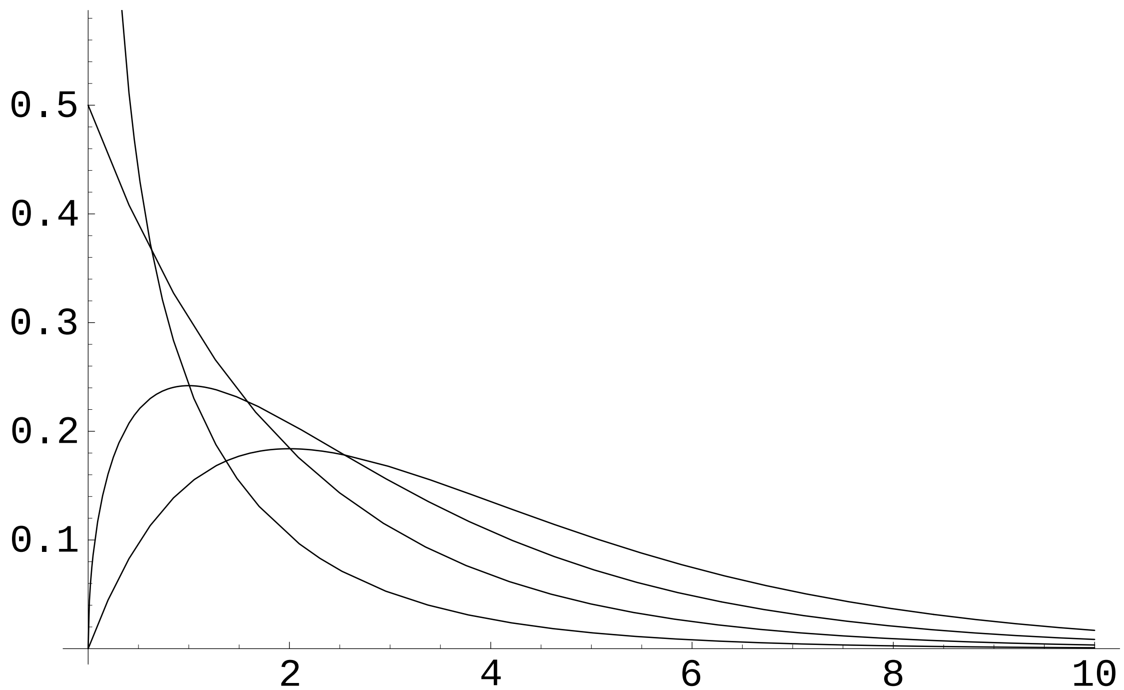
$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

(C & B p. 215)

Distributions related to normal sampling (C & B 5.3)

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



t **distribution**

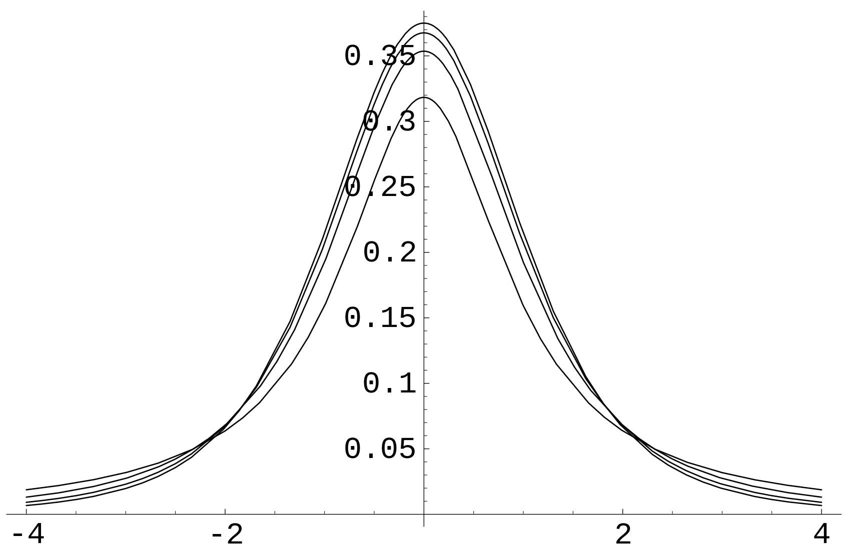
def'n: If $Z \sim N(0,1)$ independent of $X^2 \sim \chi^2_\nu$ then

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

"t with ν degrees of freedom."

Note that $E(T) = 0$ for $\nu \geq 2$ and $\text{var}(T) = \frac{\nu}{\nu - 2}$ for $\nu \geq 3$.

t_1, t_2, t_3, t_4 PDFs:



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."

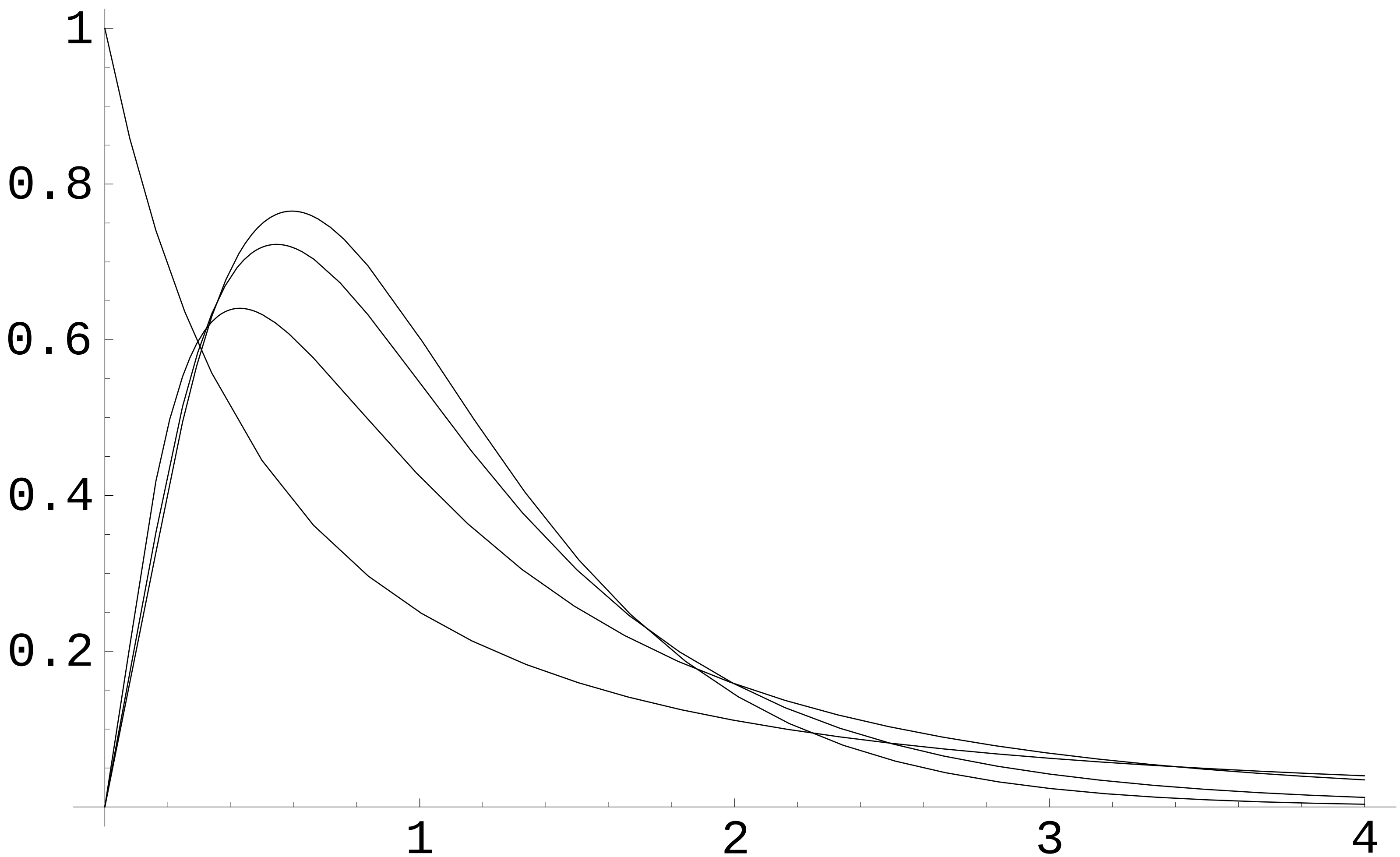
Note: The square of a t_ν random variable is an $F_{1,\nu}$ random variable. Proof:

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

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?

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



Section A.6 normal population inference

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 $H_0 : \mu = \mu_0$.

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.

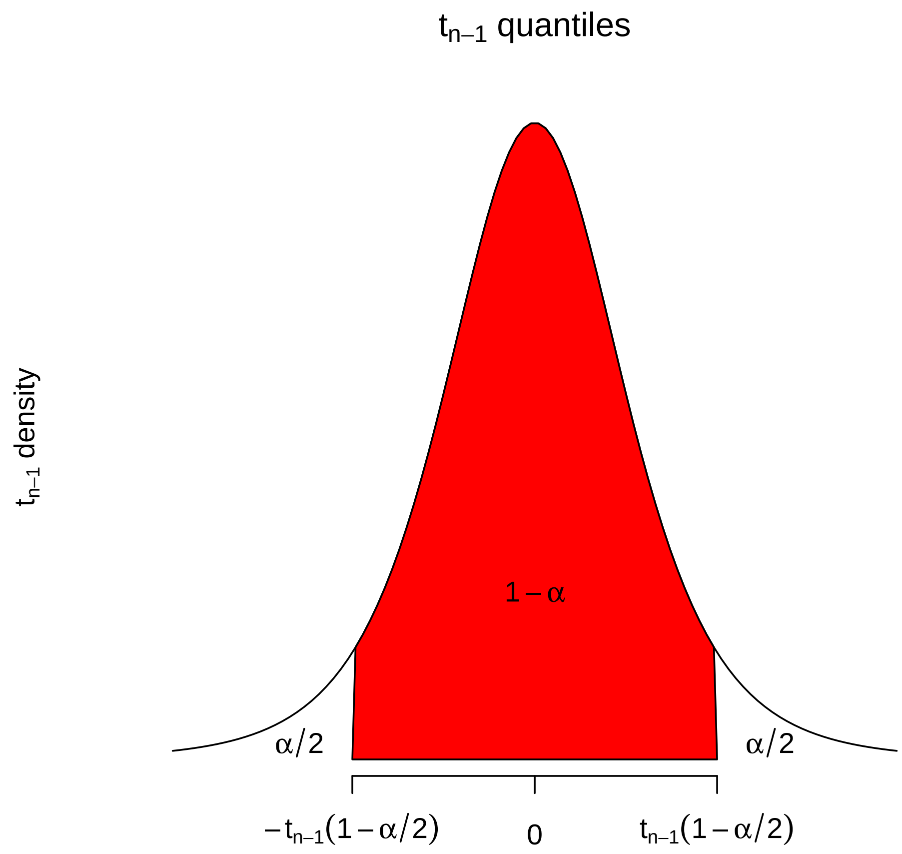
Fact: $\bar{Y}$ is independent of s^2. So then any function of $\bar{Y}$ is independent of any function of s^2.

Therefore

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

(C & B Theorem 5.3.1, p. 218)

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



Under the model

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

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

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.