# STAT 345/445 Lecture 18

## Random sample from a normal population – Section 5.3

1. Sampling from the normal distribution
   - Distribution of sample mean and sample variance
   - Correlation and independence for normals
   - The $t$-distribution
   - The $F$ distribution

# Sampling from the Normal Distribution

## Theorem 5.3.1: Distributions of $\overline{X}$ and $S^2$

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ and let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

then

(a) $\overline{X}$ and $S^2$ are independent ! ← surprising since $\bar{x}$ and $s^2$ one both functions of $x_1, \ldots x_n$.

(b) $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ ← saw this already

(c) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

# Proof of Theorem 5.3.1 (a)

(a) $\overline{X}$ and $S^2$ are independent

- Note first that $S^2$ is "over determined" with $n+1$ terms

$$X_1, X_2, \ldots, X_n, \overline{X}$$

Can write $S^2$ without $X_1$:

$$S^2 = \frac{1}{n-1}\left(\left(\sum_{i=2}^{n}(X_i - \overline{X})\right)^2 + \sum_{i=2}^{n}(X_i - \overline{X})^2\right)$$

$y_1 = \bar{x}$, $y_2\ y_3 \cdots y_n$ a function of $x_2-\bar{x}$, $x_3-\bar{x}$, $\ldots$ $x_n-\bar{x}$ (only n-1 terms)

$$(n-1)S^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 = (x_1-\bar{x})^2 + \sum_{i=2}^{n}(x_i-\bar{x})^2$$

$= A$

note: $n \cdot \bar{x} = x_1 + x_2 \cdots x_n \Rightarrow x_1 = n\bar{x} - (x_2 + x_3 \cdots x_n)$

$$= (\bar{x} - x_2) + (\bar{x} - x_3) \cdots + (\bar{x} - x_n) + \bar{x}$$

$$A = \left(\bar{x} + \sum_{i=2}^{n} (\bar{x} - x_i) - \bar{x}\right)^2$$

$$= \bar{x} + \sum_{i=2}^{n} (\bar{x} - x_i)$$

$$= \left(\sum_{i=2}^{n} (\bar{x} - x_i)\right)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

# Proof of Theorem 5.3.1 (a) - continued

- We have written $S^2$ as a function of $n-1$ terms:

$$(X_2 - \overline{X}), (X_3 - \overline{X}), \ldots, (X_n - \overline{X})$$

- Show that $\overline{X}$ and the random vector $(X_2 - \overline{X}, X_3 - \overline{X}, \ldots, X_n - \overline{X})$ are independent
  - then we have shown that $\overline{X}$ and $S^2$ are independent
- Define an $n$ dimensional transformation:

$$Y_1 = \overline{X}, \quad Y_2 = X_2 - \overline{X}, \quad Y_3 = X_3 - \overline{X}, \ \ldots, \ Y_n = X_n - \overline{X}$$

- Want to show that

$$f(y_1, y_2, y_3, \ldots, y_n) = g(y_1)\, h(y_2, y_3, \ldots, y_n)$$

for some functions $g(\cdot)$ and $h(\cdot)$

# Proof of Theorem 5.3.1 (a) - continued

- Find inverse functions $h_i(\mathbf{y})$ and Jacobian and then

$$f(\mathbf{y}) = f_{\mathbf{X}}(h_1(\mathbf{y}), \ldots, h_n(\mathbf{y}) \, |J|$$

- Assuming $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$ we have

$$f_{\mathbf{X}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-x_i^2/2) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right)$$

- Therefore

$f(\mathbf{y}) =$ 

$$y_1 = \bar{x}$$

$$y_2 = x_2 - \bar{x} \quad \Rightarrow \quad x_2 = y_2 + \bar{x} = y_2 + y_1$$

$$\vdots \qquad\qquad \vdots$$

$$y_n = x_n - \bar{x} \qquad\qquad x_n = y_n + y_1$$

and $x_1 = n\bar{x} - (x_2 + x_3 + \cdots + x_n)$

$$= n y_1 - \left[ (y_2 + y_1) + (y_3 + y_1) + \cdots (y_n + y_1) \right]$$

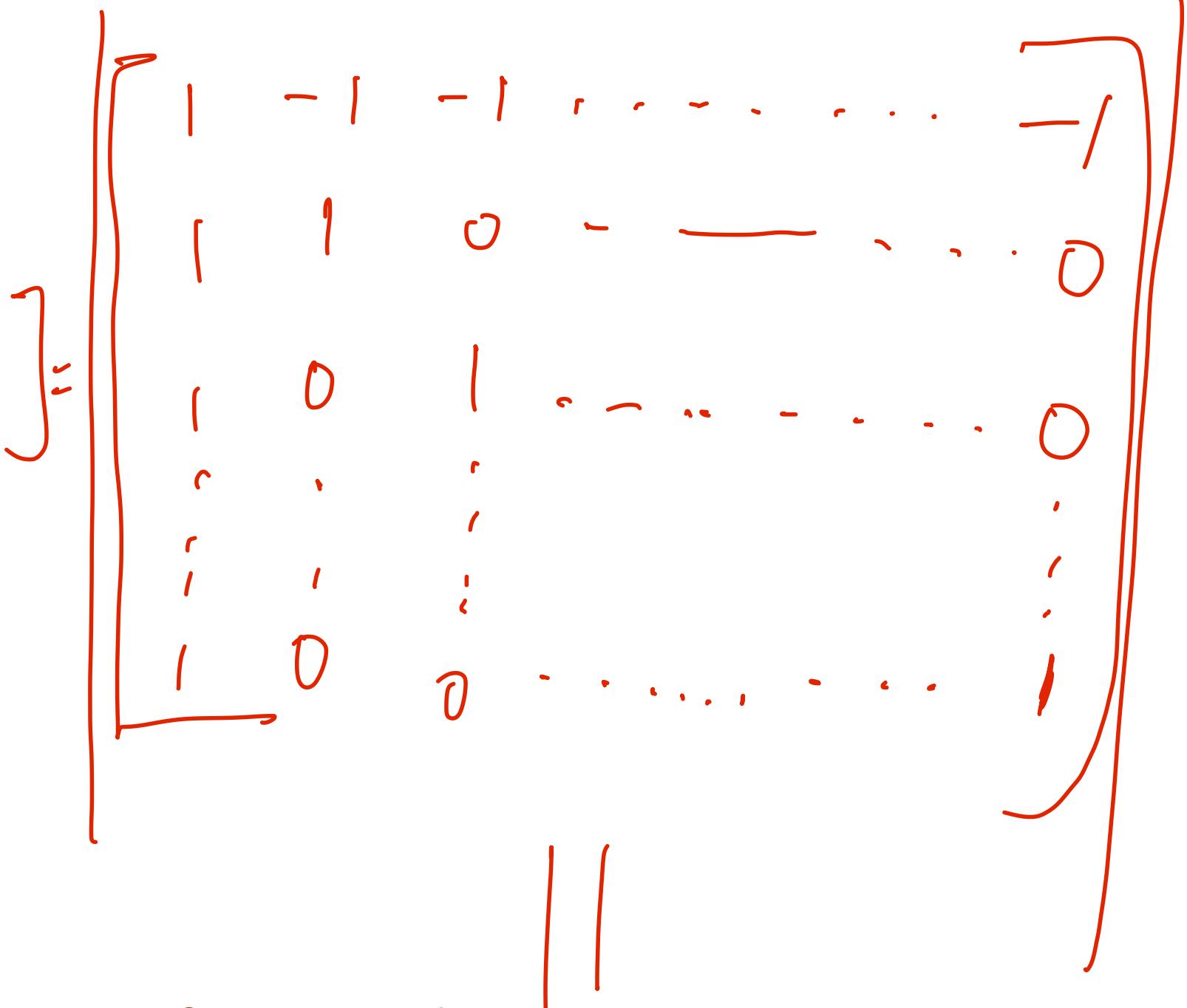
$$= y_1 - y_2 - y_3 - \cdots - y_n$$

$$\frac{\partial x_1}{\partial y_1} = 1, \quad \frac{\partial x_1}{\partial y_2} = -1, \quad \cdots, \quad \frac{\partial x_1}{\partial y_n} = -1$$

$$\frac{\partial x_2}{\partial y_1} = 1, \quad \frac{\partial x_2}{\partial y_2} = 1, \quad \frac{\partial x_2}{\partial y_3} = 0 \cdots, \quad \frac{\partial x_2}{\partial y_n} = 0$$

For $k = 2, \ldots, n$

$$\frac{\partial x_k}{\partial y_1} = 1; \cdots \frac{\partial x_k}{\partial y_k} = 1, \text{ otherwise } \frac{\partial x_k}{\partial y_j} = 0, \; j > k$$



Gauss-Elimination: add all rows 2 to n to the 1st row

$$\left| \begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \right|$$

def of a diagonal matrix = multiple of diagonal values

$$= n \rfloor_{n-1} = n \cdot \left| \begin{matrix} 1 & \cdots \\ 1 & \cdots \end{matrix} \right| = n$$

$$(*) = \int (y_1, \dots, y_n) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\underbrace{(y_1 - y_2 - \dots - y_n)^2}_{\sum_{i=2}^{n} y_i} - \frac{1}{2}\sum_{i=2}^{n}(y_i - y_1)\right) \cdot n$$

$$= n(2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\left(y_1^2 + \left(\sum_{i=2}^{n} y_i\right)^2 - 2y_1 \sum_{i=2}^{n} y_i + \dots\right)\right.$$

$$= \dots = g(y_1)\, h(y_2, \dots, y_n)$$ see slide 4

# About the $\chi^2_p$ distribution

(*1 or can show using mgf

= Gamma $\left(\frac{\sum_{i=1}^{n} p_i}{2}, 2\right)$

= $\chi^2_{p_1+\cdots+p_n}$.

## Lemma 5.3.2

(a) If $Z \sim \mathrm{N}(0,1)$ then $Z^2 \sim \chi^2_1$

(b) If $V_1, \ldots, V_n$ are independent and $V_i \sim \chi^2_{p_i}$ then

$$V_1 + \cdots + V_n \sim \chi^2_{p_1+\cdots+p_n}$$

(a): Seen before, just a univariate transformation of $Z \sim N(0,1)$

(b): Remember: $\chi^2_p$ distribution. = Gamma $(\frac{p}{2}, 2)$ distr.

Seen before: (*1) Sum of independent Gammas with same beta $\beta$

is a Gamma with that same $\beta$ and $\alpha$

= sum of the original $\alpha$'s $\Rightarrow \gamma_1 + \dots + \gamma_n \sim$

$\text{Gamma}\left(\sum_{i=1}^{n} \frac{p_i}{2}, 2\right)$

Note: if $x_1, \dots, x_n$ are iid $N(\mu, \sigma^2)$,

then $\frac{x_i - \mu}{\sigma} \overset{iid}{\sim} N(0,1)$

$$\Rightarrow \frac{(x_i-\mu)^2}{\sigma^2} \sim \chi^2_1$$

$$\Rightarrow \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} \sim \chi^2_n$$

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

not $\mu$

but $(n-1)S^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2$

"loose one degree of freedom"

# Proof of Theorem 5.3.1 (c)

(c) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$X_1, X_2, \ldots, X_n$ iid $N(0,1)$

- Proof by induction. Set

$$\overline{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i \quad \text{and} \quad S_k^2 = \frac{1}{k-1}\sum_{i=1}^{k}(X_i - \overline{X}_k)^2$$

and note that

recursive formulas

$$\overline{X}_{k+1} = \frac{k\overline{X}_k + X_{k+1}}{k+1}$$

and $$kS_{k+1}^2 = (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2$$

Can show on your own

- Start with $k = 2$ ...

$X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, show $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

$k=2$ show that $\frac{S_2^2}{\sigma^2} \sim \chi^2_1$

$$\frac{S_2^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{2}(X_i - \bar{X}_2)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{2}\left[\left(X_1 - \frac{X_1+X_2}{2}\right)^2 + \left(X_2 - \frac{X_1+X_2}{2}\right)^2\right]$$

$$= \frac{1}{\sigma^2}\left(\left(\frac{X_1-X_2}{2}\right)^2 + \left(\frac{X_2-X_1}{2}\right)^2\right) = \frac{2}{\sigma^2}\left(\frac{X_1-X_2}{2}\right)^2$$

$$= \frac{(X_1-X_2)^2}{2\sigma^2}$$

$X_1 - X_2 \sim N(0, 2\sigma^2) \Rightarrow \frac{X_1-X_2}{\sqrt{2\sigma^2}} \sim N(0,1)$

$\Rightarrow \frac{(X_1-X_2)^2}{2\sigma^2} \sim \chi^2_1$

Assume that $\frac{(k-1)S_k^2}{\sigma^2} \sim \chi^2_{k-1}$ and show that

$\frac{kS_{k+1}^2}{\sigma^2} \sim \chi^2_k$

$$\frac{kS_{k+1}^2}{\sigma^2} = \frac{k-1}{\sigma^2}S_k^2 + \frac{k}{\sigma^2(k+1)}(X_{k+1} - \bar{X}_k)^2$$

$\overset{?}{\sim} \chi^2_1$ indep to first term

$\sim \chi^2_{k-1}$ by Assume

by the two $*$ $\rightarrow \chi^2_{k-1+1} = \chi^2_k$

Independent: $S_k^2$ and $\bar{X}_k$ are indep. (Theorem)

$S_k^2$ and $\bar{X}_{k+1}$ are indep.

and $X_{k+1}$ and $\bar{X}_k$ are indep $\Rightarrow$

Two terms are indep (X1 i.i.d) and

$$X_{k+1} - \bar{X}_k \sim N(0, \sigma^2 + \sigma^2/k)$$

$N(\mu, \sigma^2)$ $\quad$ $N(\mu, \sigma^2/k)$ $\quad$ $\frac{k\sigma^2 + \sigma^2}{k} = \frac{k+1}{k}\sigma^2$

$$\Rightarrow \frac{X_{k+1} - \bar{X}_k}{\sqrt{\frac{k+1}{k}\sigma^2}} \sim N(0,1) \Rightarrow \frac{(X_{k+1} - \bar{X}_k)^2}{\frac{k+1}{k}\sigma^2} \sim \chi_1^2$$

$$\Rightarrow \frac{k}{(k+1)\sigma^2}(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2 \ *$$

# Correlation and independence

- The normal distribution has a very special property

**Independence $\Leftrightarrow$ correlation = 0**

Let $(X_1, X_2)$ be a bivariate normal random vector. Then $X_1$ and $X_2$ are independent if and only if $\rho = \mathrm{Cor}(X_1, X_2) = 0$

Recall the pdf of a bivariate normal pdf:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

# The student's $t$ distribution

- Students-$t$ distribution was developed because we want to know the distribution of the following statistic

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

  - Used in inference about the mean
- Introduced by W. S. Gosset, who wrote under the alias "Student"
- The legend:
  - Gosset derived the $t_m$ distribution while working for the Guinness Brewery in Dublin. In fear of competition he was forbidden to publish his analysis of brewery data and hence he wrote under the pseudonym Student.

# The student's $t$ distribution

## Definition: The $t$ distribution

Let $U$ and $V$ be independent random variables and $U \sim N(0,1)$ and $V \sim \chi^2_p$. The the distribution of

$$T = \frac{U}{\sqrt{V/p}}$$

is called the **$t$ distribution with $p$ degrees of freedom, or $t_p$**

What is the pdf for $T$?

# Deriving the pdf for the student's $t$ distribution

## Let's do the derivation that Gosset did 120 years ago!

- Let $U \sim N(0,1)$ and $V \sim \chi^2_p$ be independent
- Want the pdf of

$$T = \frac{U}{\sqrt{V/p}}$$

- Strategy: Do a bi-variate transformation

$$(U, V) \mapsto (T, W)$$

  and then find the marginal of $T$

- Set $W = V$

# Deriving the pdf for the student's $t$ distribution - cont

# The student's $t$ distribution

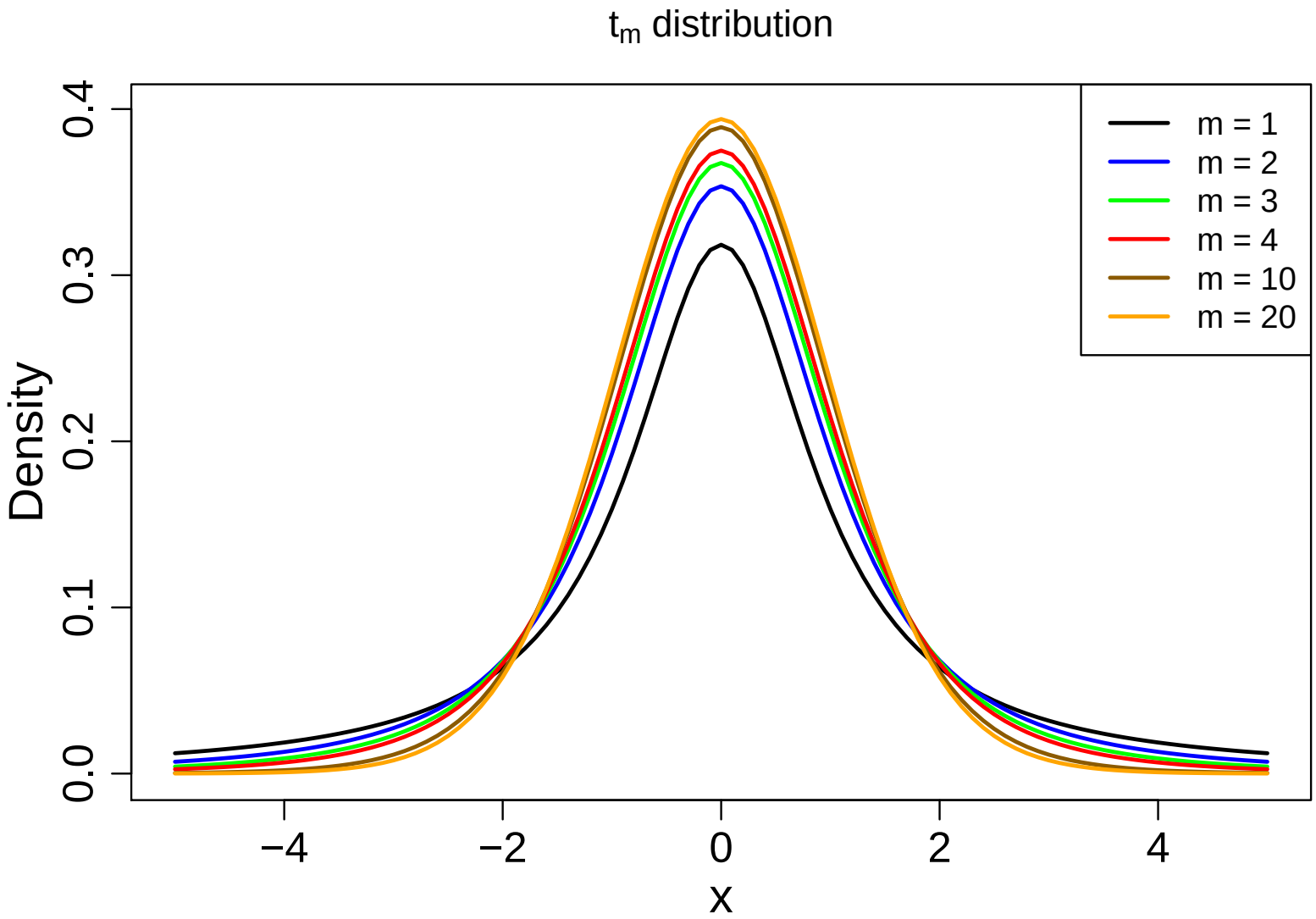
- The pdf of the $t_p$ distribution is

$$f(x) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{\left(1 + x^2/p\right)^{(p+1)/2}}, \quad -\infty < x < \infty$$

The parameter $p$ ("degrees of freedom") is an integer

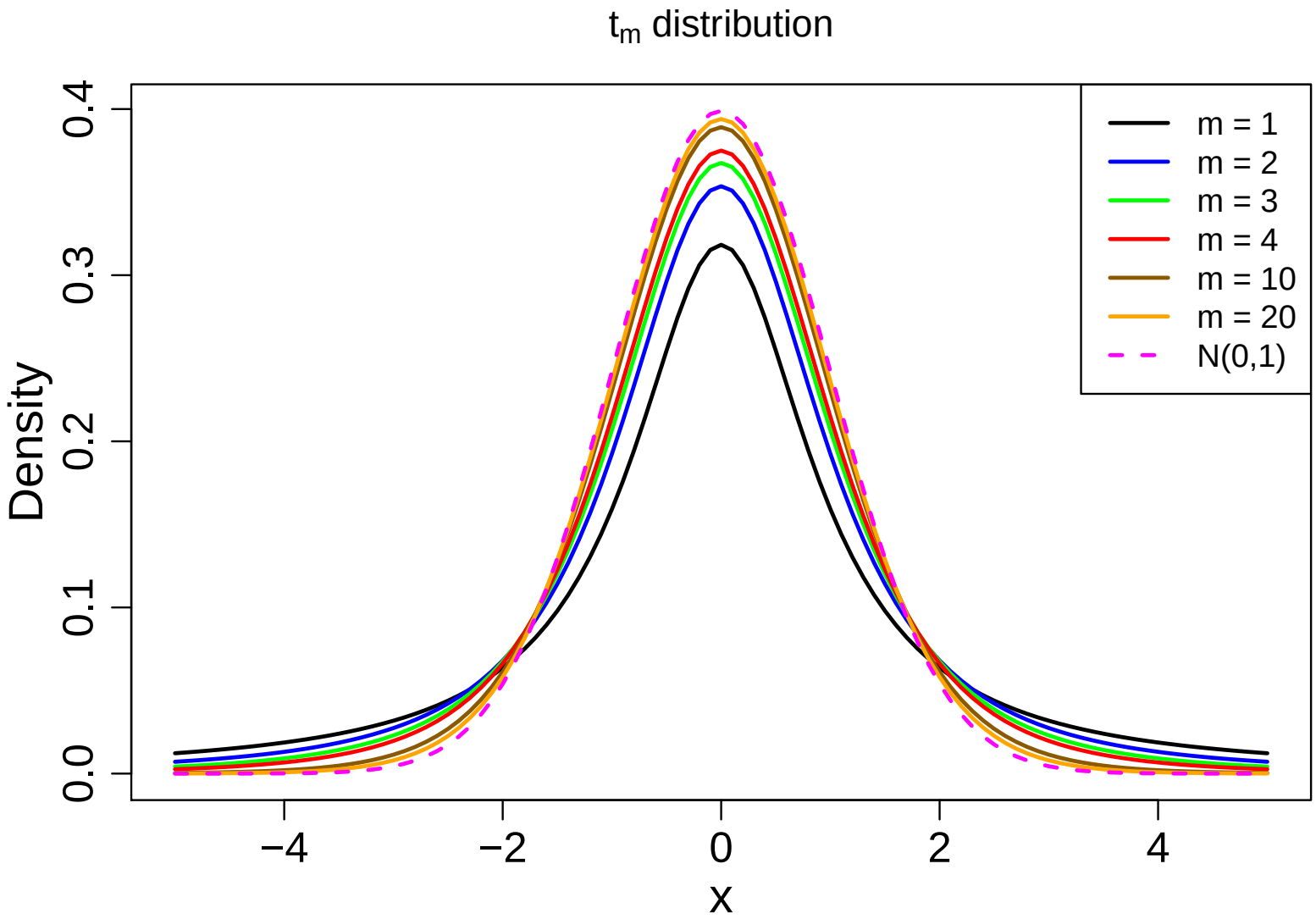
# The $t_p$ pdfs and the standard normal pdf

## As $p \to \infty$ the $t_p$ approaches $N(0, 1)$



# The $t_p$ pdfs and the standard normal pdf

## As $p \to \infty$ the $t_p$ approaches $N(0,1)$



# The $T$ statistic

## The $T$ statistic

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$. Then the statistic

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

has **$t$ distribution with $n-1$ degrees of freedom, or $t_{n-1}$**

proof ...

$$\Rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)s^2}{\sigma^2(n-1)}} \sim t_{n-1}$$

$$= \frac{(\bar{x}-\mu)\sqrt{n}}{s} \Rightarrow \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

# Snedecor's $F$ distribution[1]

- Comparing two variances: $S_1^2/S_2^2$
- Linear models: Ratio of sums-of-squares

## Def: $F_{u,\nu}$-distribution

Let $X \sim \chi^2_p$ and $Y \sim \chi^2_q$ be independent random variables. The distribution of

$$U = \frac{X/p}{Y/q} \sim F_{p,q}$$

is called the *F distribution with p and q degrees of freedom*

[1] *Trivia:* George W. Snedecor founded the first academic department of statistics in the United States, at Iowa State University in 1947.

If distribution $X \sim \chi^2_p$, $Y \sim \chi^2_q$ $U = \frac{X/p}{Y/q}$

Define $V = Y$ (for example)

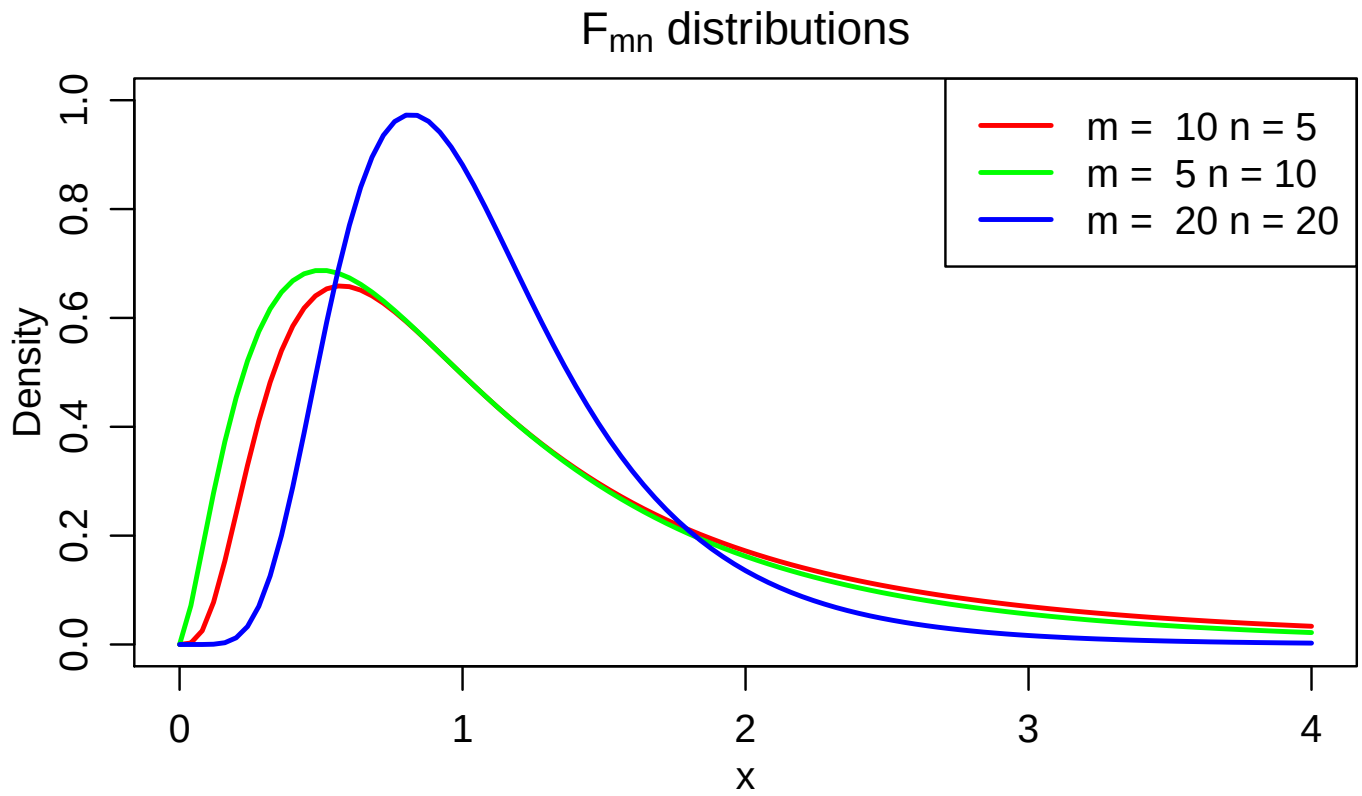
$f(x,y) = f(x) \cdot f(y)$ find $f(u,v)$ and

$f(u) = \int f(u,v)\,dv$

# F-distribution –pdf

## $F$-distribution - pdf

$$f(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{x^{(p/2)-1}}{(1+(p/q)x)^{(p+q)/2}}, \quad 0 < x < \infty$$



# The $F$ statistic

## The $F$ statistic

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean $\mu_X$ and variance $\sigma_X^2$.
- Let $Y_1, Y_2, \ldots, Y_m$ be a random sample from a normal distribution with mean $\mu_Y$ and variance $\sigma_Y^2$.

Then the statistic

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}$$

has **$F$ distribution with $n-1$ and $m-$ degrees of freedom, or** $F_{n-1,m-1}$

# $F$-distribution – properties

Some useful (univariate) transformations of the $F$ distribution

## Theorem 5.3.8

(a) If $X \sim F_{p,q}$ then $1/X \sim F_{q,p}$ ⇒ tables in their text books only know upper quantiles

(b) If $X \sim t_q$ then $X^2 \sim F_{1,q}$

(c) If $X \sim F_{p,q}$ then $\frac{(p/q)X}{1+(p/q)X} \sim \text{Beta}(q/2, p/2)$

⇒ In STAT 325: Marginal t-test in the Summary table is equivalent to the last F-test in the Sequential Anova table.

Once $f_{1-\alpha/2,p,q} = \frac{1}{f_{\alpha/2,q,p}}$

# $T$ and $F$ statistics – uses

- The $T$ statistic is used in inference of the mean of a normal distribution when variance is unknown, e.g.
  - Population mean of one or two populations
    - $t$-test, two-sample $t$-test, paired $t$-test ...
  - Regression coefficients in a normal linear regression model
- The $F$ statistics is used in many situations, e.g.
  - to test for equality of variances from two independent populations
  - Analysis of Variance (ANOVA)
  - comparing nested models in normal linear regression

# Linear model example

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.6100262  7.8418734   0.205    0.837
PercPoverty      4.0099847  0.2807055  14.285  < 2e-16 ***
PercUnemployed  -2.2020717  0.4689496  -4.696 3.57e-06 ***
IncomePerCapita  0.0018748  0.0003045   6.157 1.69e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.55 on 433 degrees of freedom
Multiple R-squared:  0.3246,    Adjusted R-squared:  0.3199
F-statistic: 69.36 on 3 and 433 DF,  p-value: < 2.2e-16

> anova(Fit)
Analysis of Variance Table

Response: CrimeRate
                Df Sum Sq Mean Sq F value    Pr(>F)
PercPoverty      1  60376   60376 142.989 < 2.2e-16 ***
PercUnemployed   1  11476   11476  27.179 2.878e-07 ***
IncomePerCapita  1  16008   16008  37.912 1.690e-09 ***
Residuals      433 182832     422
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```