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

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

surprising since $\overline{X}$ and S^2 are both functions of $X_1, \ldots, X_n$.

then

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

(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

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

showed on the white-board

Can write S^2 without X_1:

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

= a function of

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

(only n-1 terms)

Proof of Theorem 5.3.1 (a) - continued

- We have written S^2 as a function of $n-1$ terms: $Y_2, \ldots, Y_n$

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

- 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

$(X_1, \ldots, X_n) \longmapsto (Y_1, \ldots, Y_n)$ Show: Y_1 and $(Y_2, \ldots, Y_n)$ are indep.

- Define an n dimensional transformation:

$$Y_1 = \overline{X}, \quad \underbrace{Y_2 = X_2 - \overline{X}, \quad Y_3 = X_3 - \overline{X}, \quad \ldots, \quad Y_n = X_n - \overline{X}}_{S^2 \text{ is a func. of these.}}$$

- 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 $h(\cdot)$

We did not have time in class to finish all the details, so I added them afterwards:

Finding the inverse functions:

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

n-dim transformation

$$\Rightarrow \quad X_2 = Y_2 + \bar{X} = Y_2 + Y_1$$

$$X_3 = Y_3 + \bar{X} = Y_3 + Y_1$$

$$\vdots$$

$$X_n = Y_n + Y_n$$

$$x_1 = nY_1 - x_2 - x_3 - \cdots - X_n = nY_1 \underbrace{- (Y_2 - Y_1) - \cdots - (Y_3 - Y_1)}_{\substack{n-1 \text{ terms} \\ \Rightarrow -(n-1)Y_1}}$$

$$= Y_1 - Y_2 - Y_3 - \cdots - Y_n$$

$$\frac{\delta x_1}{\delta y_1} = 1, \quad \frac{\delta x_1}{\delta y_2} = -1, \ldots, \quad \frac{\delta x_1}{\delta y_n} = -1$$

$$\frac{\delta x_2}{\delta y_1} = 1, \quad \frac{\delta x_2}{\delta y_2} = 1, \quad \frac{\delta x_2}{\delta y_3} = 0, \ldots, \quad \frac{\delta x_2}{\delta y_n} = 0$$

... for $k = 2, \ldots, n$ we get $\frac{\delta x_k}{\delta y_1} = 1$ and $\frac{\delta x_k}{\delta y_k} = 1$

other derivatives are zero

The Jacobian is therefore:

$$\left| \begin{bmatrix} \frac{\delta x_1}{\delta y_1} & \cdots & \frac{\delta x_1}{\delta y_n} \\ \vdots & & \\ \frac{\delta x_n}{\delta y_1} & \cdots & \frac{\delta x_n}{\delta y_n} \end{bmatrix} \right| = \left| \begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \right|$$

determinant

Apply Gauss-Elimination (does not change determinant)

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

added rows $2, \dots, n$ to row 1

* determinant of a triangular matrix is the product of the diagonal values

$$\Rightarrow |J| = |n| = n$$

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

$$f(\mathbf{y}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\left(\left(y_1 - \sum_{k=2}^{n} y_k\right)^2 + \sum_{i=2}^{n} (y_1 + y_i)^2\right)\right) \cdot n$$

if $= g(y_1)\, h(y_2, \cdots, y_n)$ the proof is done $\rightarrow$

$$f(y) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\left(\left(y_1 - \sum_{k=2}^{n} y_k\right)^2 + \sum_{i=2}^{n} (y_1 + y_i)^2\right)\right) \cdot n$$

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

cancel

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

$$= \underbrace{n(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\left(n y_1^2\right)\right)}_{= g(y_1)} \underbrace{\exp\left(-\frac{1}{2}\left(\sum_{k=2}^{n} y_k + \sum_{i=2}^{n} y_i^2 \right)\right)}_{= h(y_2, y_3, \dots, y_n)}$$

$\Rightarrow$ Y_1 and $(Y_2, Y_3, \dots, Y_n)$ are independent

$\Rightarrow$ $\overline{X}$ and S^2 are independent.

About the χ^2_p distribution

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

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

*or can show using mgfs.

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: χ^2_p distr. $= \text{Gamma}\left(\frac{p}{2}, 2\right)$ distr.

seen before*: Sum of indep. Gammas with same β and is a Gamma with that same β and α = sum of the original α's

$\Rightarrow V_1 + \cdots + V_n \sim \text{Gamma}\left(\sum_{i=1}^n \frac{p_i}{2}, 2\right)$

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

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

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

can show on your own.

- Start with $k = 2$ ...

Note: If $X_1, \dots, X_n$ are iid $N(\mu, \sigma^2)$

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

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

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

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

not μ

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

"loose one deg. of freedom"

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

$$T = \frac{\overset{N(0,1)}{U}}{\sqrt{V/p}} \quad \sim \quad \sqrt{\chi^2_p/p}$$

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

$(U, V) \longmapsto (T, W) \qquad T = \frac{U}{\sqrt{V/p}}$ and $W = V$

$\Rightarrow \quad V = W = h_1(t, w)$ and

$u = t\sqrt{V/p} = t\sqrt{w/p} = h_2(t, w)$

$\frac{\delta u}{\delta t} = \sqrt{\frac{w}{p}} \quad , \quad \frac{\delta u}{\delta w} = \frac{t}{\sqrt{p}} \frac{1}{2\sqrt{w}} = \frac{t}{2\sqrt{pw}} \; , \; \frac{\delta v}{\delta t} = 0 \; , \; \frac{\delta v}{\delta w} = 1$

$$\Rightarrow J = \left| \begin{bmatrix} \sqrt{\frac{w}{p}} & \frac{t}{2\sqrt{pw}} \\ 0 & 1 \end{bmatrix} \right| = \sqrt{\frac{w}{p}}$$

$$f_{T,W}(t, w) = f_{U,V}(h_1(t,w), h_2(t,w))\,|J|$$

U, V independent

$$\Rightarrow f_{U,V}(u,v) = f_U(u)\, f_V(v)$$

Remember: $\chi^2_p = \text{Gamma}\left(\frac{p}{2}, 2\right)$

$$= \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \; \frac{1}{\Gamma(p/2)\, 2^{p/2}} v^{\frac{p}{2}-1} e^{-v/2}$$

for $-\infty < u < \infty$ and $0 < v < \infty$

$$\Rightarrow f_{T,W}(t,w) = \frac{1}{\sqrt{2\pi}\, 2^{p/2}\, \Gamma(p/2)} w^{p/2-1} \exp\left(-\tfrac{1}{2} t^2 w/p - w/2\right) \cdot \sqrt{\frac{w}{p}}$$

for $-\infty < t < \infty$ and $0 < w < \infty$

$$\Rightarrow f_T(t) = \int_0^\infty \frac{1}{\sqrt{\pi}\, 2^{\frac{p+1}{2}}\, \Gamma(p/2) \sqrt{p}} w^{\overbrace{\frac{p-1}{2}}^{=\alpha-1}} \exp\Big(-w \overbrace{\left(\frac{t^2}{2p} + \frac{1}{2}\right)}^{=\frac{1}{\beta}}\Big)\, dw$$

set $\alpha = \frac{p-1}{2} + 1 = \frac{p+1}{2}$ and $\beta = \frac{1}{\frac{t^2+p}{2p}} = \frac{2p}{t^2+p} = \frac{2}{t^2/p + 1}$

$$= \frac{1}{\sqrt{\pi}\, 2^{\frac{p+1}{2}}\, \Gamma(p/2) \sqrt{p}} \; \Gamma(\alpha)\beta^\alpha \underbrace{\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} e^{-w/\beta}\, dw}_{=1}$$

$$\Rightarrow \quad f_T(t) = \frac{1}{\sqrt{\pi}\; 2^{\frac{p+1}{2}}\; \Gamma(p/2)\sqrt{p}} \qquad \Gamma(\alpha)\beta^{\alpha}$$

$$= \frac{1}{\sqrt{\pi}\; 2^{\frac{p+1}{2}}\; \Gamma(p/2)\sqrt{p}}\; \Gamma\left(\frac{p+1}{2}\right)\left(\frac{2}{t^2/p+1}\right)^{\frac{p+1}{2}}$$

$$= \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \cdot \frac{1}{\sqrt{\pi p}}\; \frac{1}{\left(t^2/p+1\right)^{\frac{p+1}{2}}} \qquad \text{for } -\infty < t < \infty$$

The student's t distribution

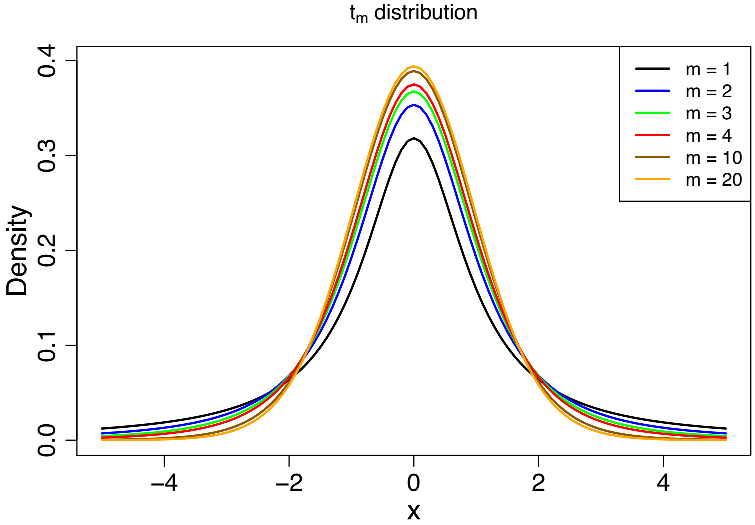
- The pdf of the t_p distribution is

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

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

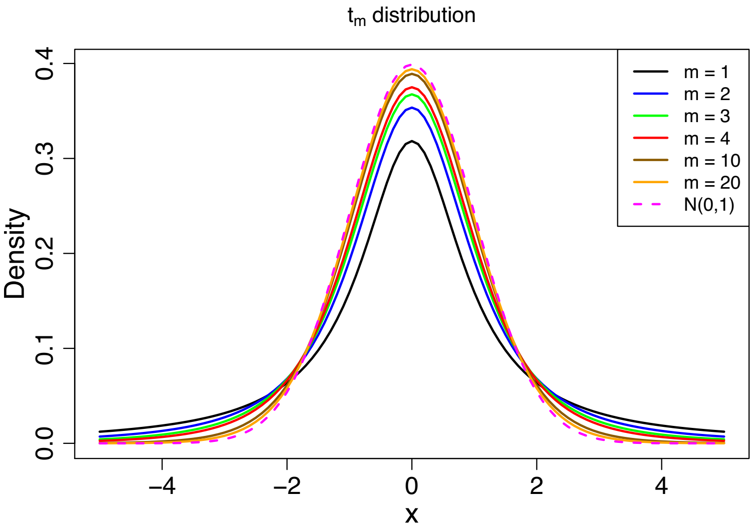
The t_p pdfs and the standard normal pdf

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 μ and variance σ^2. Then the statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \quad \sim t_{n-1}$$

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

proof ...

Know: $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$\overline{X}$ and S^2 are independent $\Rightarrow$ $\frac{\overline{X}-\mu}{\sigma}$ and $\frac{(n-1)S^2}{\sigma^2}$ are indep.

$$\Rightarrow \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\quad}$$

Know: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$\bar{X}$ and S^2 are independent $\Rightarrow$ $\dfrac{\bar{X}-\mu}{\sigma}$ and $\dfrac{(n-1)S^2}{\sigma^2}$ are indep.

$$\Rightarrow \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} \sim t_{n-1}$$

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

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 $X \sim \chi_p^2$ and $Y \sim \chi_q^2$ be independent random variables. The distribution of

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

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.

F-distribution

$X \sim \chi^2_p$, $Y \sim \chi^2_q$ $\qquad u = \dfrac{X/p}{Y/q}$

Define $V = Y$ (for example)

$f(x,y) = f(x) f(y)$

Find $f(u,v)$ and

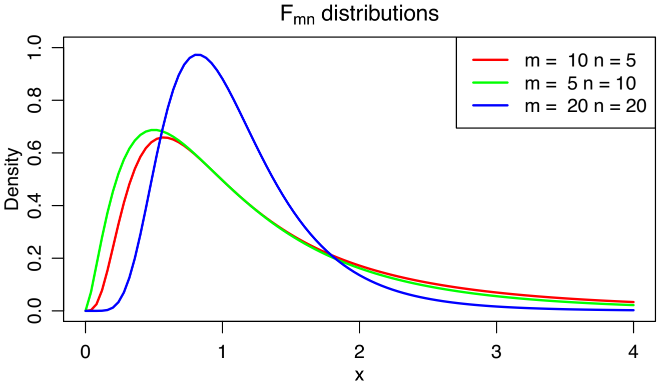
$f(u) = \int f(u,v)\, dv$

$\rightarrow$ A good exercize!

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

$\frac{(n_x-1)S_x^2}{\sigma_x^2} \sim \chi^2_{n_x-1}$ $\qquad$ $\frac{(n_y-1)S_y^2}{\sigma_y^2} \sim \chi^2_{n_y-1}$

The F statistic

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ_X and variance σ_X^2.
- Let $Y_1, Y_2, \ldots, Y_m$ be a random sample from a normal distribution with mean μ_Y and variance σ_Y^2.

Then the statistic

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

$n_x = n$
$n_y = m$

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

$$\Rightarrow \quad \frac{(n_x-1)S_x^2/\sigma_x^2(n_x-1)}{(n_y-1)S_y^2/\sigma_y^2(n_y-1)} \sim F_{n_x-1,\,n_y-1}$$

F-distribution – properties

Some useful (univariate) transformations of the F distribution

Theorem 5.3.8

(a) If $X \sim F_{p,q}$ then $1/X \sim F_{q,p}$

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

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

=> Tables in stat text books only show upper quantiles since $f_{1-\alpha/2, p, q} = \frac{1}{f_{\alpha/2, q, p}}$

→ In STAT 325: Marginal t-test in the summary table is equivalent to the last F-test in the (sequential) ANOVA table.

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

t-tests for

summary table

$H_0: \beta_k = 0$ vs. $H_1: \beta_k \neq 0$

```
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
```

$= \sqrt{37.912}$

same test