STAT 515 -- Chapter 6: Sampling Distributions

Definition: <u>Parameter</u> = a number that characterizes a population (example: population mean μ) – it's typically <u>unknown</u>.

<u>Statistic</u> = a number that characterizes a sample

(example: sample mean $\bar{X}$) – we can calculate it from our sample data.

We use the sample mean $\bar{X}$ to estimate the population mean μ.

Suppose we take a sample and calculate $\bar{X}$.

Probably not.

Will $\bar{X}$ equal μ? Will $\bar{X}$ be close to μ? We hope so.

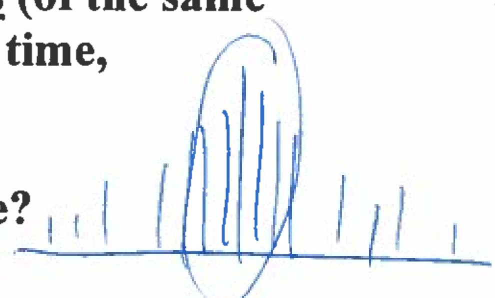
Suppose we take another sample and get another $\bar{X}$.

Will it be same as first $\bar{X}$? Will it be close to first $\bar{X}$? We hope so (Probably, depends on sample size)

Probably not.

- **What if we took <u>many repeated samples</u> (of the same size) from the same population, and each time, calculated the sample mean?**

What would that set of $\bar{X}$ values look like?

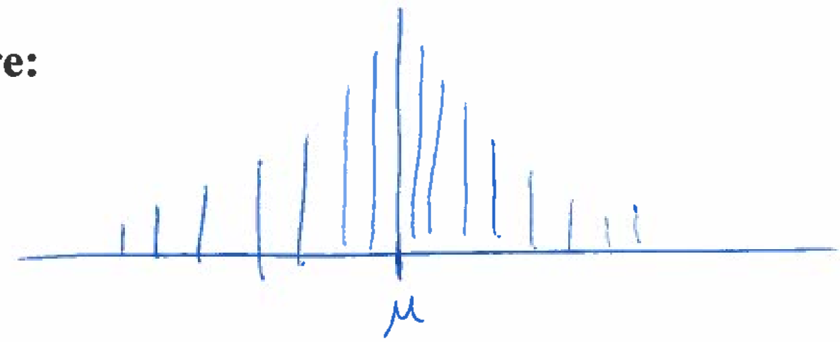


The <u>sampling distribution</u> of a statistic is the distribution of values of the statistic in all possible samples (of the same size) from the same population.

Consider the sampling distribution of the sample mean

$\bar{X}$ when we take samples of size n from a population with mean μ and variance σ^2.

Picture:



The sampling distribution of $\bar{X}$ has mean μ and standard deviation $\sigma/\sqrt{n}$.

σ is std. dev. of the original population.

Notation:

$$\mu_{\bar{X}} = \mu \iff E(\bar{X}) = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

<u>Point Estimator</u>: A statistic which is a single number meant to estimate a parameter.

It would be nice if the average value of the estimator (over repeated sampling) equaled the target parameter.

An estimator is called <u>unbiased</u> if the mean of its sampling distribution is equal to the parameter being estimated.

Examples: $E(\bar{X}) = \mu$, so $\bar{X}$ is an unbiased estimator of μ.

$E(s^2) = \sigma^2$, so s^2 is an unbiased estimator of σ^2.

$E(s) \neq \sigma$, so s is a biased estimator of σ.

Another nice property of an estimator: we want the spread of its sampling distribution to be as small as possible.

The standard deviation of a statistic's sampling distribution is called the standard error of the statistic.

The standard error of the sample mean $\bar{X}$ is $\sigma/\sqrt{n}$.

Note: As the sample size gets larger, the spread of the sampling distribution gets smaller.

When the sample size is large, the sample mean varies less across samples. $\leftarrow$ good

Evaluating an estimator:

(1) Is it unbiased?

(2) Does it have a small standard error?

Central Limit Theorem

We have determined the center and the spread of the sampling distribution of $\bar{X}$. What is the shape of its sampling distribution?

Case I: If the distribution of the original data is normal, the sampling distribution of $\bar{X}$ is normal. (This is true no matter what the sample size is.)

Case II: Central Limit Theorem: If we take a random sample (of size *n*) from any population with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is approximately normal, if the sample size is large.

$$\bar{X} \overset{\cdot}{\sim} N(\underbrace{\mu}_{\text{mean}}, \underbrace{\sigma/\sqrt{n}}_{\text{std. error}}) \quad \text{if } n \text{ is large}$$

How large does *n* have to be?
Our rule of thumb: If $n \geq 30$, we can apply the CLT result.

Pictures: See R code when sampling from an exponential distribution.

As *n* gets larger, the closer the sampling distribution looks to a normal distribution.

<u>Why is the CLT important</u>? Because when $\bar{X}$ is (approximately) normally distributed, we can answer probability questions about the sample mean.

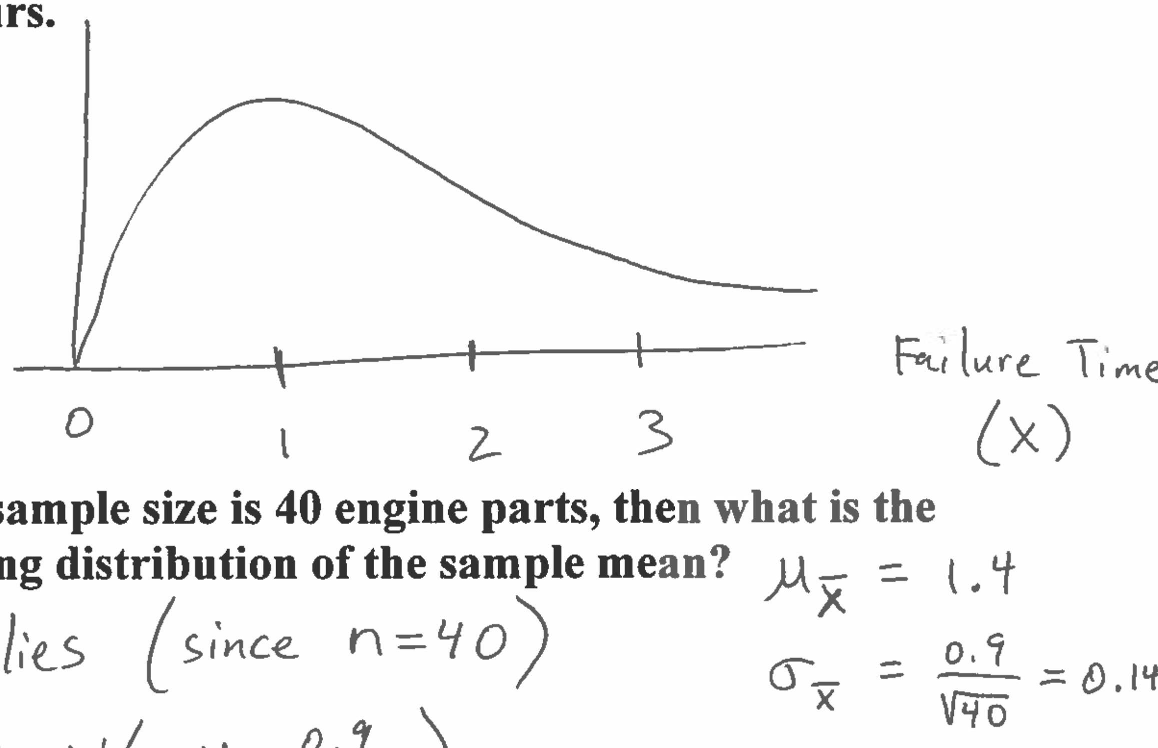
Standardizing values of $\bar{X}$:

If $\bar{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n}$, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution.

Example: Suppose we're studying the failure time (at high stress) of a certain engine part. The failure times have a mean of 1.4 hours and a standard deviation of 0.9 hours.



If our sample size is 40 engine parts, then what is the sampling distribution of the sample mean?

$\mu_{\bar{X}} = 1.4$

$\sigma_{\bar{X}} = \frac{0.9}{\sqrt{40}} = 0.1423$

CLT applies (since $n=40$)

$\bar{X} \dot{\sim} N\left(1.4, \frac{0.9}{\sqrt{40}}\right)$

$\Rightarrow \bar{X} \dot{\sim} N(1.4, 0.1423)$

What is the probability that the sample mean will be greater than 1.5?

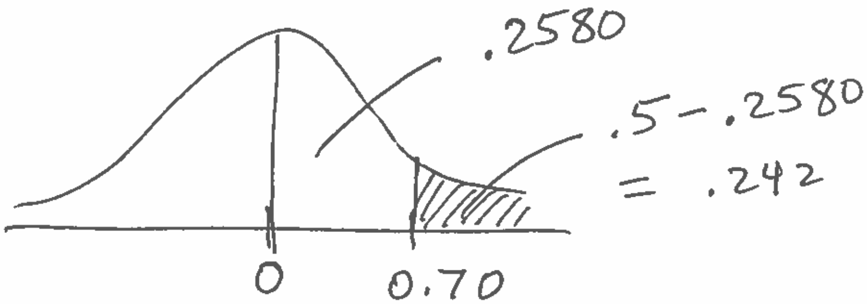
$P(\bar{X} > 1.5) \approx P(Z > 0.70)$

$= \boxed{.242}$

$\bar{X} = 1.5$

$\Rightarrow Z = \dfrac{1.5 - 1.4}{0.9/\sqrt{40}}$

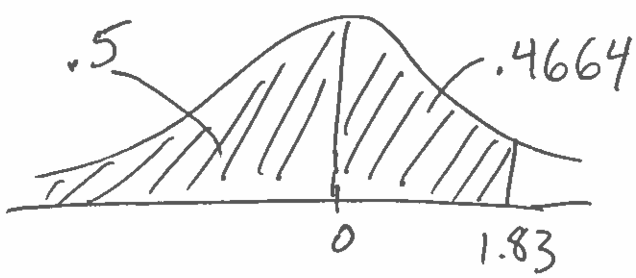
$= \dfrac{1.5 - 1.4}{0.1423} = 0.70$



Example: Suppose lawyers' salaries have a mean of \$90,000 and a standard deviation of \$30,000 (highly skewed). Given a sample of lawyers, can we find the probability the sample mean is less than \$100,000 if *n* = 5? No **If *n* = 30?** Yes, we can use CLT.

$P(\bar{X} < 100{,}000) \approx P(Z < 1.83) = \boxed{.9664}$

$\bar{X} = 100{,}000 \Rightarrow Z = \dfrac{100{,}000 - 90{,}000}{30000/\sqrt{30}} = 1.83$



Other Sampling Distributions

In practice, the population standard deviation σ is typically unknown.

We estimate σ with *s*.

But the quantity $\dfrac{\overline{X}-\mu}{s/\sqrt{n}}$ **no longer has a standard normal distribution.**

Its sampling distribution is as follows:

• If the data come from a normal population, then the statistic $T=\dfrac{\overline{X}-\mu}{s/\sqrt{n}}$ **has a t-distribution ("Student's t") with *n* – 1 degrees of freedom (the parameter of the t-distribution).**

• The t-distribution resembles the standard normal (symmetric, mound-shaped, centered at zero) but it is more spread out.
• The fewer the degrees of freedom, the more spread out the t-distribution is.
• As the d.f. increase, the t-distribution gets closer to the standard normal.

Picture:

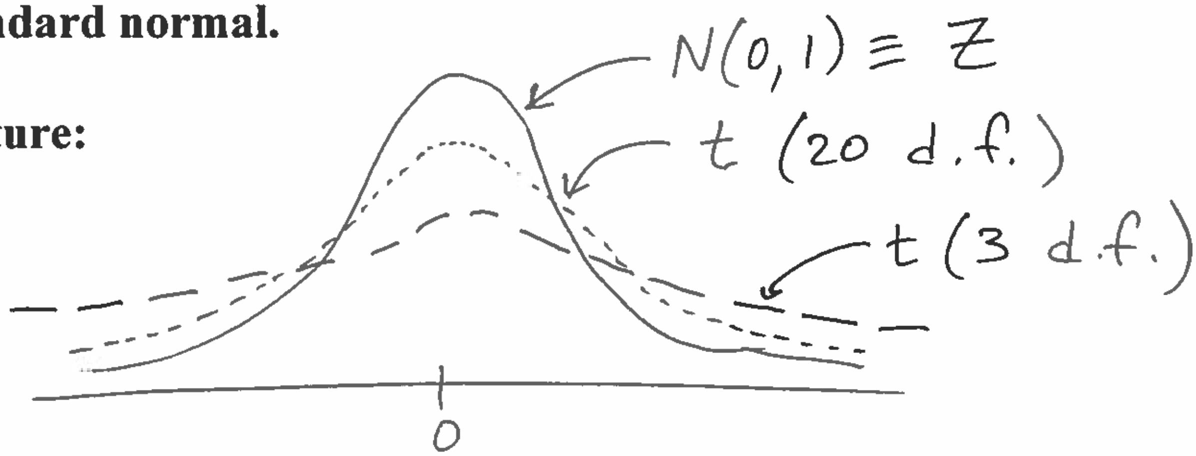
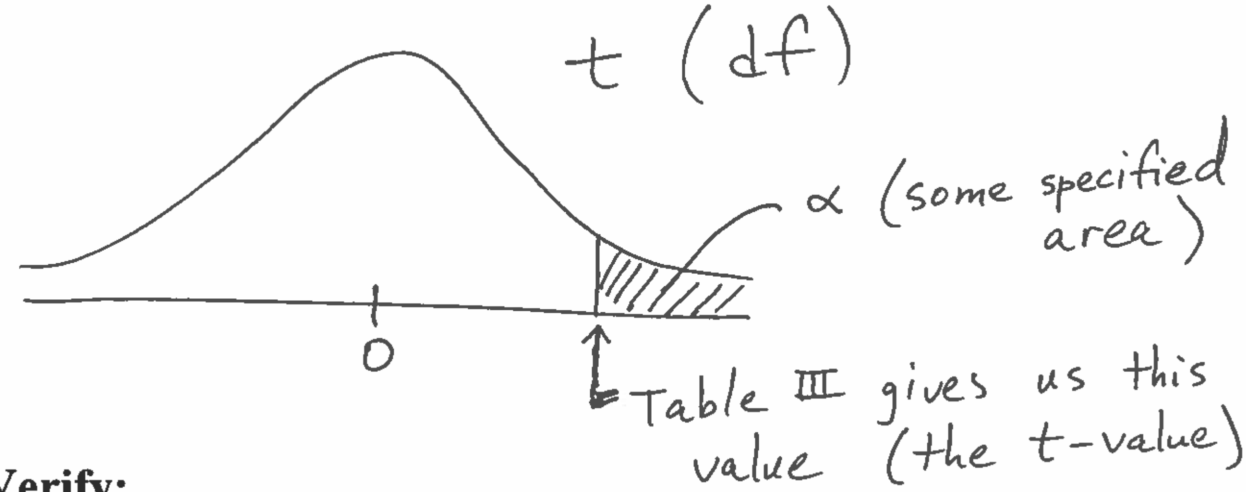


Table ~~VI~~ III gives values of the t-distribution with specific areas to the right of these values:



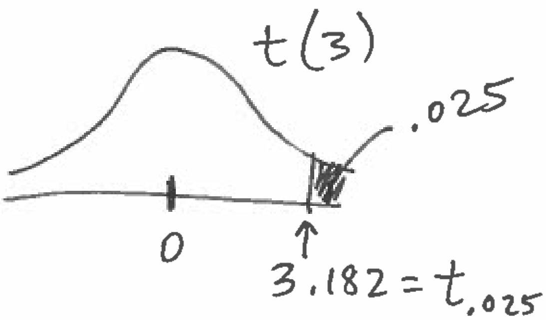
Verify:

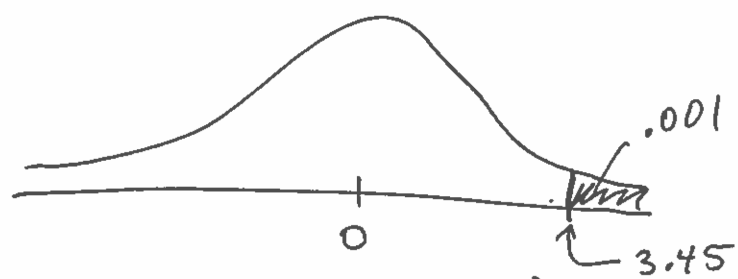
In t-distribution with 3 d.f., area to the right of 3.182 **is .025. (Notation: For 3 d.f., $t_{.025}$ =** 3.182 **)**

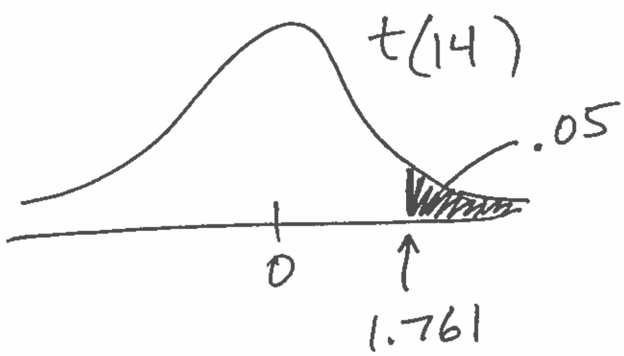
In t with 14 d.f., area to the right of 1.761 **is .05.**

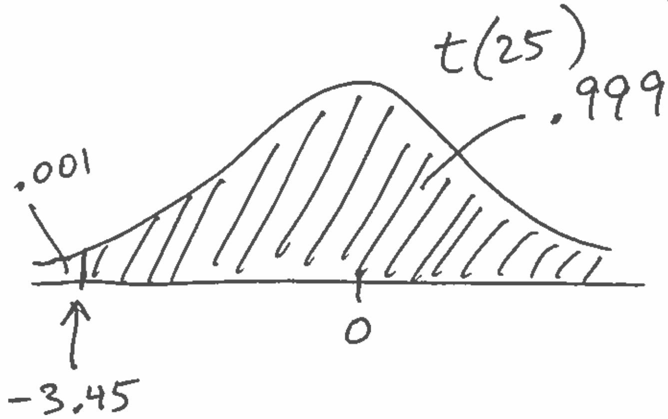
For 14 d.f., $t_{.05} = 1.761$

In t with 25 d.f., area to the right of _______ is .999.









The χ^2 (Chi-square) Distribution

Suppose our sample (of size n) comes from a normal population with mean μ and standard deviation σ.

Then $\dfrac{(n-1)s^2}{\sigma^2}$ has a χ^2 distribution with $n-1$ degrees of freedom.

- **The χ^2 distribution takes on positive values.**
- **It is skewed to the right.**
- **It is less skewed for higher degrees of freedom.**
- **The mean of a χ^2 distribution with $n-1$ degrees of freedom is $n-1$ and the variance is $2(n-1)$.**

<u>Fact</u>: If we add the squares of n independent standard normal r.v.'s, the resulting sum has a χ^2_n distribution.

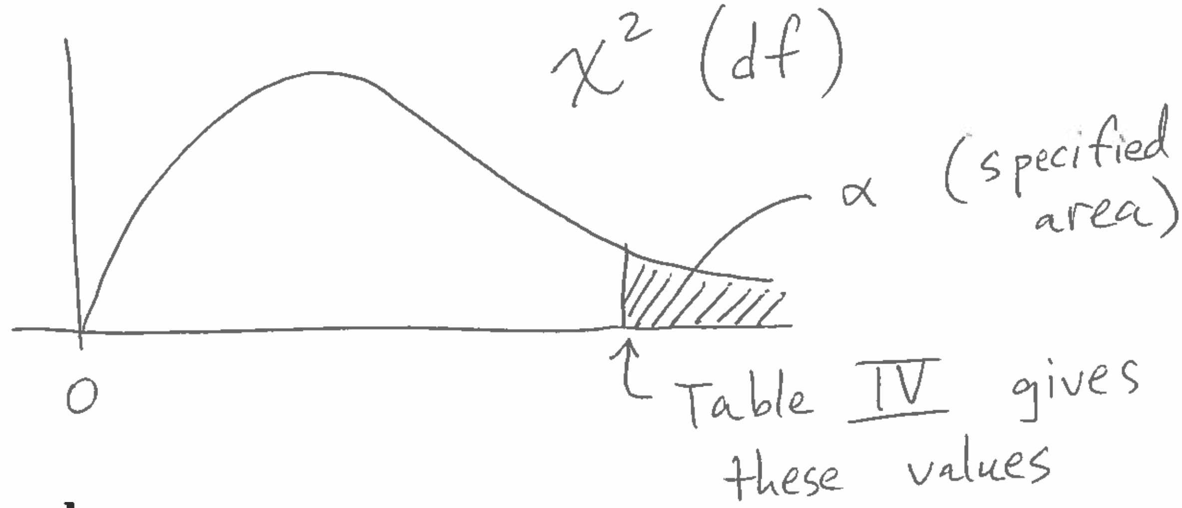
Note that $\dfrac{(n-1)s^2}{\sigma^2} = \dfrac{n-1}{\sigma^2} \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$$

- If we had used μ rather than $\bar{X}$, this would be n independent standard normal r.v.'s, squared and added up $\rightarrow$ Would have a χ^2_n distribution.

We sacrifice one d.f. by estimating μ with $\bar{X}$, so it is χ^2_{n-1}.

Table ~~VII~~ IV gives values of a χ^2 r.v. with specific areas to the right of those values.

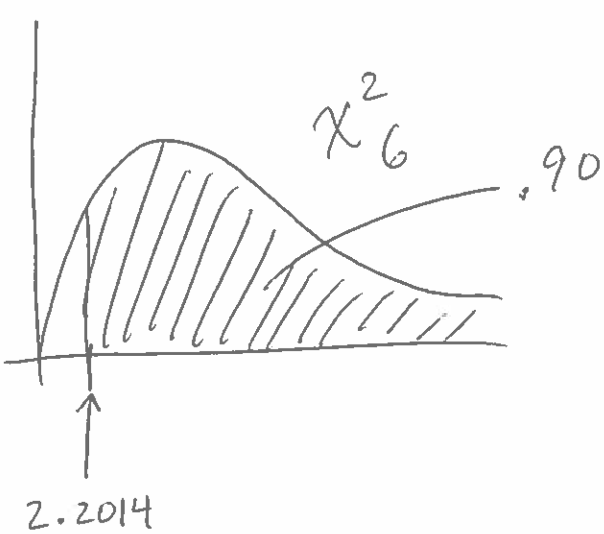


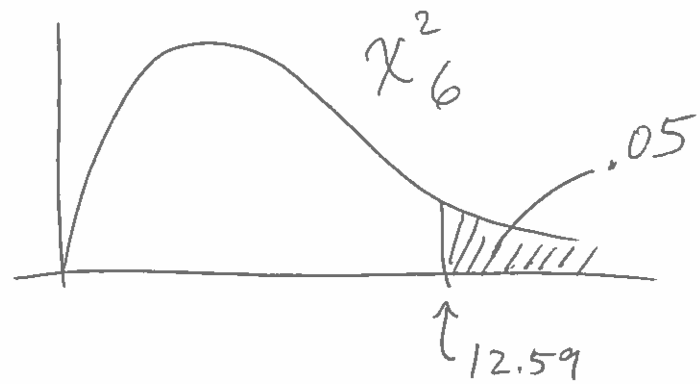
Examples:

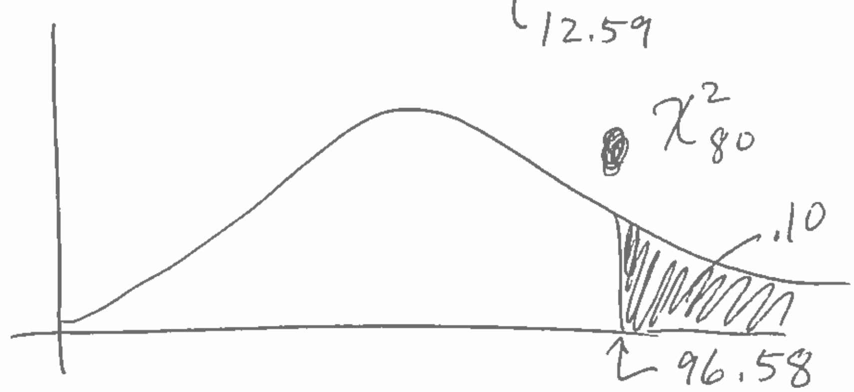
For χ^2 with 6 d.f., area to the right of 2.204 **is .90.**

For χ^2 with 6 d.f., area to the right of 12.59 **is .05.**

For χ^2 with 80 d.f., area to the right of 96.58 **is .10.**







The F Distribution

The quantity $\dfrac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}$ where the two χ^2 r.v.'s are independent, has an F-distribution with $n_1 - 1$ "numerator degrees of freedom" and $n_2 - 1$ denominator degrees of freedom.

So, if we have independent samples (of sizes n_1 and n_2) from <u>two</u> normal populations, note:

$$\frac{\dfrac{\cancel{(n_1-1)}\, S_1^2}{\sigma_1^2\, \cancel{(n_1-1)}}}{\dfrac{\cancel{(n_2-1)}\, S_2^2}{\sigma_2^2\, \cancel{(n_2-1)}}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F-distribution with $(n_1 - 1, n_2 - 1)$ d.f.

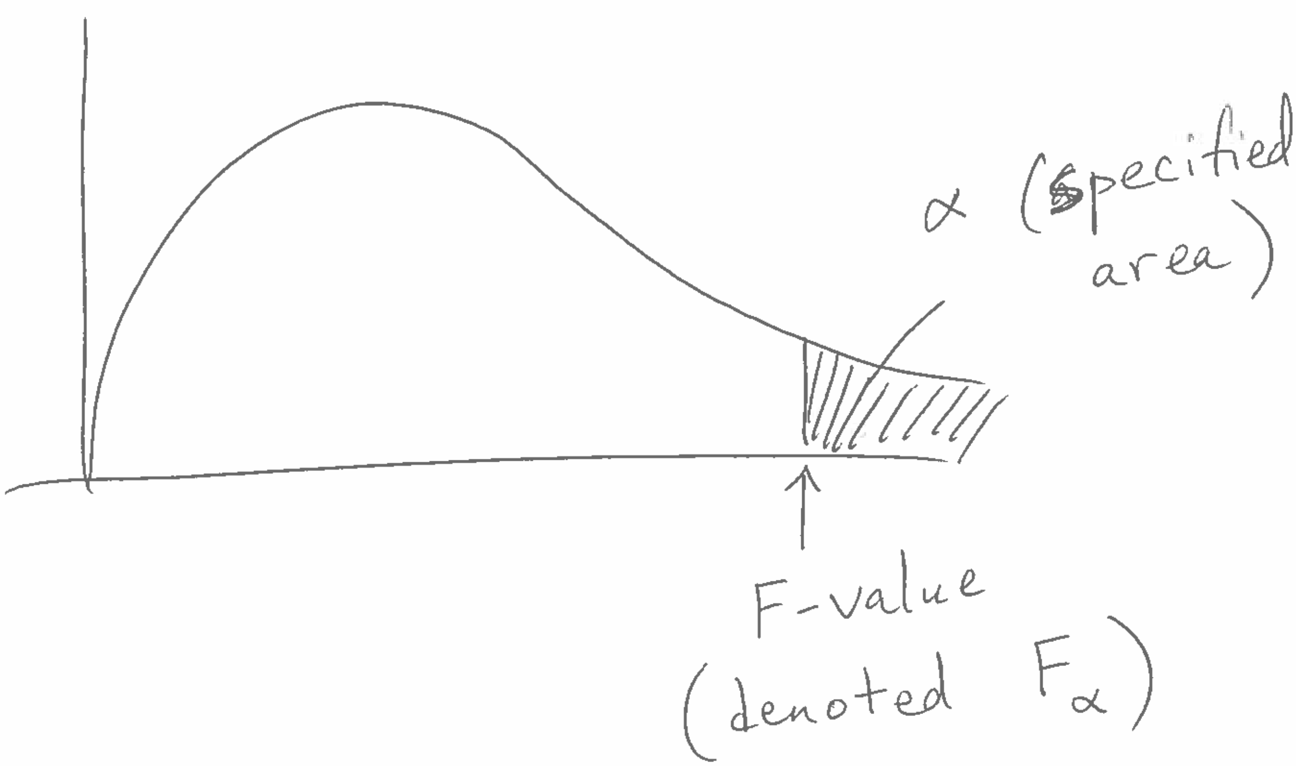


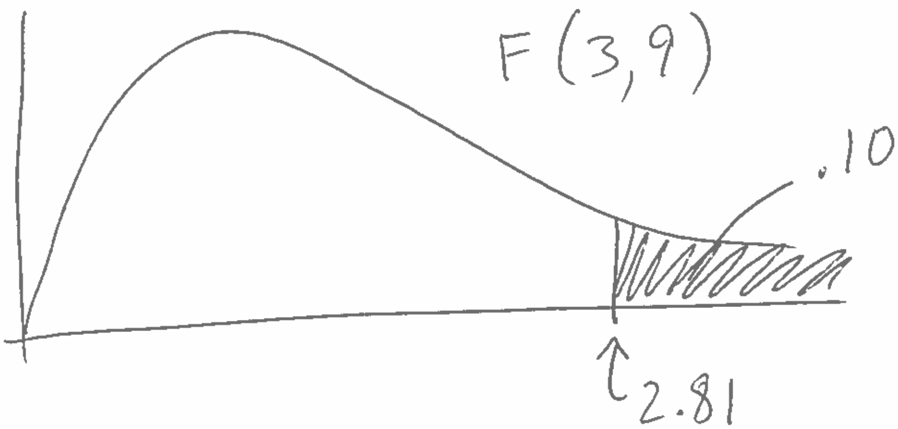
Table ~~VIII~~ V gives values of F r.v. with area .10 to the right.
Table ~~IX~~ VI gives values of F r.v. with area .05 to the right.
Table ~~X~~ VII gives values of F r.v. with area .025 to the right.
Table ~~XI~~ VIII gives values of F r.v. with area .01 to the right.

Verify:

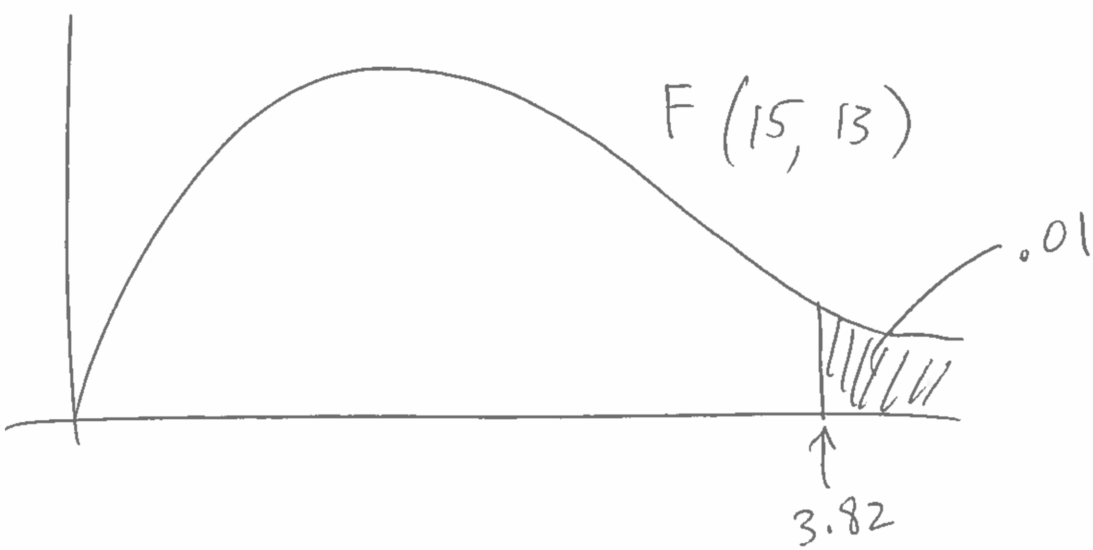
For F with (3, 9) d.f., 2.81 has area 0.10 to right.

For F with (15, 13) d.f., 3.82 has area 0.01 to right.

• These sampling distributions will be important in many inferential procedures we will learn.



Look in Table V



Look in Table VIII