Uni-variate Statistics

- Types of Random Variables
- Description of RV's
  - using Probability Distributions
  - using Parameters
- Selected Probability Distributions

Values of Random Variables

- Quantitative, e.g., income, rainfalls
- Qualitative, e.g., color
- Semi-quantitative or semi-qualitative, e.g., course letter grade, income group

  express order not magnitude

Types of Quantitative RV's

- Discrete RV

  assume values that do not form a continuous range, e.g., number of attendants

  Its possible values are not limited to only integers.
- Continuous RV's

  assume a continue range of values, e.g., temperature

Probability Distributions

- Tools to describe the Uncertainty of a quantitative random variable
- Popular tools
  - Probability Mass Function (PMF)
  - Probability Density Function (pdf)
  - Cumulative Distribution Function (CDF)

Estimation and Statistical Inference

Why?

—Unknown real probability distribution

—Unknown parameters

—Estimation accuracy

Data Sampling

- Sampling/observing
  - —sampling schemes
  - —sample size or number of observations
- Presentation of a data sample
  - —graph
  - —Calculated statistics

Graphing data

- to show the distribution
- type of graphs, e.g.,
  - —histogram
  - —line graphs
  - —bar graphs
  - —pie chart

Calculated Statistics

- to estimate the unknown parameters
- Popular statistics
  - — sample mean
  - — sample variance
  - — sample skewedness
  - — sample kurtosis(peakedness or tail-thickness)

Random Sampling (1)

- Realization of the previous observation will not influence the results of the next observations
- Sample of size n

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

Random Sampling (2)

- X_i's are random variables having identical probability distributions.
- X_i's are independent from each other.
- Values of X_i does not influence Values of X_j for $i \neq j$

Sample Mean

- Estimator of population mean (μ_X)

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample Variance

- Estimator of population variance (σ_X^2)

$$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

- Why *n-1* not *n*?

Sample Standard Deviation

- Estimator of population SD or σ_X

$$s_X = \sqrt{s_X^2}$$

Sample Skewedness

- Estimator of population skewedness

$$S_k = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_X} \right)^3$$

- Why?

Sample Kurtosis (Peakedness)

- Estimator of population peakedness

$$K = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- Why?

Sampling Distributions

- Probability of Calculated Statistics
- Note that they are random
- Assumptions
  1) $X \sim N(\mu, \sigma^2)$
     $E(X) = \mu$
     $V(X) = \sigma^2$
  2) Random sampling

Distribution of $\overline{X}$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$E(\overline{X}) = \mu$$

$$V(\overline{X}) = \sigma^2 / n$$

Distribution of s_X^2

$$s_X^2 \sim \text{non-standard}$$

$$(n-1)\frac{s_X^2}{\sigma^2} \sim \chi^2(n-1)$$

$$E(s_X^2) = \sigma^2$$

$$V(s_X^2) = \frac{2\sigma^4}{n-1}$$

Unbiased Estimators (1)

- Definition: $\hat{\theta}$ is an unbiased estimator of θ

  if $E(\hat{\theta}) = \theta$

- Note that $\overline{X}$ is an unbiased estimator of μ.

  That is, $E(\overline{X}) = \mu$

Unbiased Estimators (2)

- s_X^2 is an unbiased estimator of σ^2
- Reason for dividing by (*n-1*) is to make the estimators unbiased.
- Same reason for the formulae of sample skewedness and peakedness

Desired Properties of an Estimator

- Unbiasedness
- Efficiency or smallest possible variance
- If cannot find such an estimator, the next best is a **consistent** estimator
  - asymptotically unbiased, i.e., bias ->0 as n-> ∞
  - asymptotically diminishing variance, i.e., variance->0 as n-> ∞

Estimator and Estimate

- Estimator is a formula to estimate a parameter
- Estimate is a calculated or observed value to be used as a proxy for the parameter

Standardized $\overline{X}$ (1)

- Subtract with its mean and divide by its standard deviation

$$Z_{Cal} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

- Note that if σ is known, Z can be calculated given μ

Standardized $\overline{X}$ (2)

- Subtract with its mean and divide by its sample standard deviation

$$t_{Cal} = \frac{\overline{X} - \mu}{s_X / \sqrt{n}} \sim t(n-1)$$

- Does not need to know σ to calculate *t*.
- Why?

Standardized $\overline{X}$ (3)

$$t_{cal} = \frac{\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)\dfrac{s_X^2}{\sigma^2}}{n-1}}} \sim t(n-1)$$

Note: see the creation of t-distributed variable

Central Limit Theorem for $\overline{X}$ (1)

- Violation of normal distribution of X
  but still E(X) = μ
  V(X) = σ^2
- $\overline{X}$ is approximately normal when n-> ∞

or $\sqrt{n}\left(\overline{X} - \mu\right)$ is asymptotically normal

$$\sqrt{n}\left(\overline{X} - \mu\right) \overset{A}{\sim} N\left(0, \sigma^2\right)$$

Central Limit Theorem for $\overline{X}$ (2)

Z_{cal} is approx. distributed as a standard normal

t_{cal} is approx. t-distributed.

Confidence Interval for μ

Interval estimator for μ (cf. Point estimator)

<u>Concept</u>

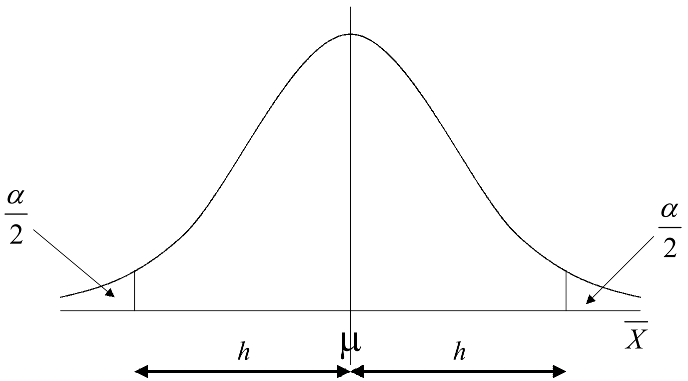
Confidence Interval for $\mu = \overline{X} \pm h$

where h = half-width of the confidence interval

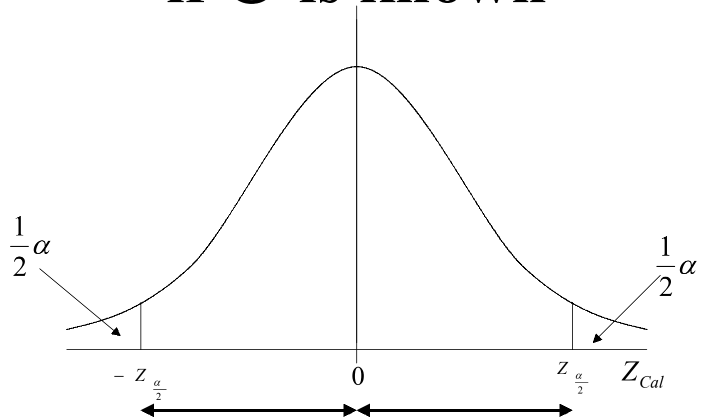
Factors Affecting *h*

- Sample size (n). h decreases as n increases.
- Confidence Level. h increases as confidence level increases
- Note that
  - Confidence Level = $(1-\alpha)100\%$ where α is the significance level (chosen by the one who estimates)
  - Standard confidence levels are 90%, 95% and 99%
  - Standard values of α are 0.10, 0.05 and 0.01

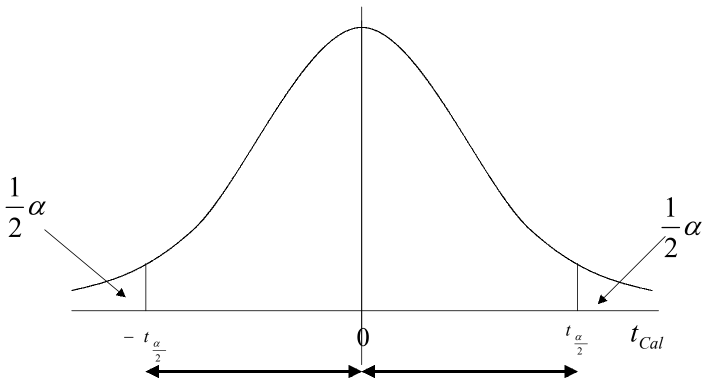
How to choose *h* (1)



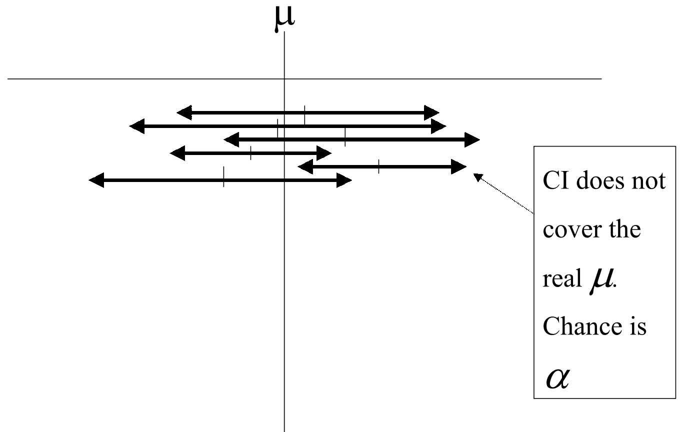
How to choose *h* (2) if σ is known



How to choose *h* (3) if σ is unknown



Interpretation of CI



Confidence Interval for σ^2

Interval estimator for σ^2 (cf. Point estimator)

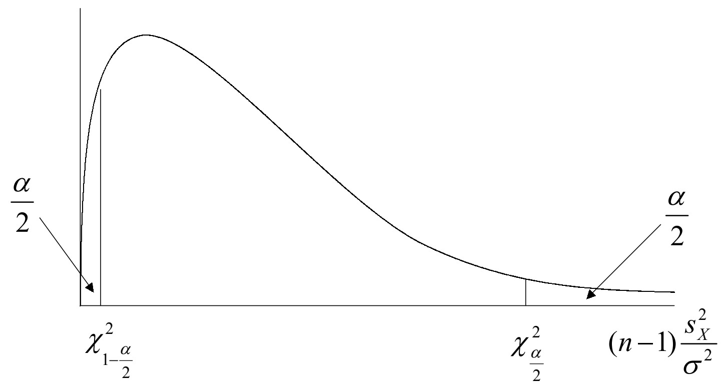
<u>Define</u>

Confidence Interval for σ^2 = $\{L,U\}$

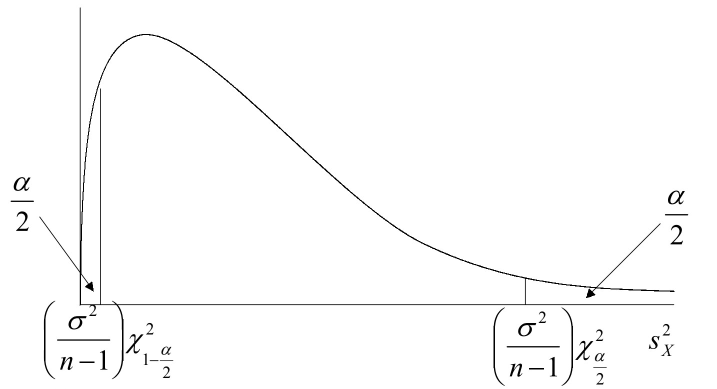
where L = lower limit of the confidence interval

U = upper limit of the confidence interval

How to choose L,U (1)



How to choose L,U (2)



How to choose *L,U* (3)

Note that

$$\left(\frac{\sigma^2}{n-1}\right)\chi^2_{1-\frac{\alpha}{2}} < s_X^2 < \left(\frac{\sigma^2}{n-1}\right)\chi^2_{\frac{\alpha}{2}}$$

Take inverse

$$\left(\frac{n-1}{\sigma^2}\right)\frac{1}{\chi^2_{1-\frac{\alpha}{2}}} > \frac{1}{s_X^2} > \left(\frac{n-1}{\sigma^2}\right)\frac{1}{\chi^2_{\frac{\alpha}{2}}}$$

How to choose *L,U* (4)

Multiply with $\sigma^2 s_X^2$

$$(n-1)\frac{s_X^2}{\chi^2_{1-\frac{\alpha}{2}}} > \sigma^2 > (n-1)\frac{s_X^2}{\chi^2_{\frac{\alpha}{2}}}$$

Flip over

$$(n-1)\frac{s_X^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < (n-1)\frac{s_X^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

How to choose *L,U* (5)

(1-α)100% Confidence Interval for

$$\sigma^2 = \left\{(n-1)\frac{s_X^2}{\chi^2_{\frac{\alpha}{2}}}, (n-1)\frac{s_X^2}{\chi^2_{1-\frac{\alpha}{2}}}\right\}$$

One-sided Confidence Interval for σ^2

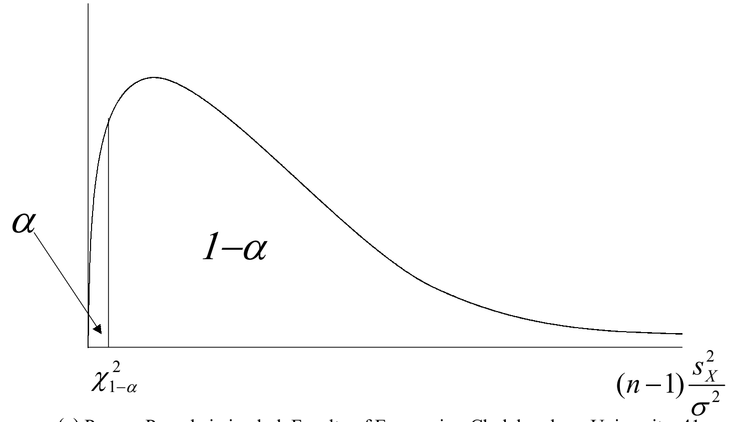
Concept

Note that $\sigma^2 \geq 0$ always.

In general, we are more concerned about the upper side of the limit.

One-sided Confidence Interval for σ^2 = $\{0,U\}$

How to choose *U* for One-sided CI (1)



How to choose *U* for One-sided CI (2)

For (1-α)100% chance that

$$\chi^2_{1-\alpha} < (n-1)\frac{s_X^2}{\sigma^2}$$

$$0 < \sigma^2 < (n-1)\frac{s_X^2}{\chi^2_{1-\alpha}}$$

How to choose *U* for One-sided CI (3)

(1-α)100% One-sided Confidence Interval for

$$\sigma^2 = \left\{0, (n-1)\frac{s_X^2}{\chi^2_{1-\alpha}}\right\}$$

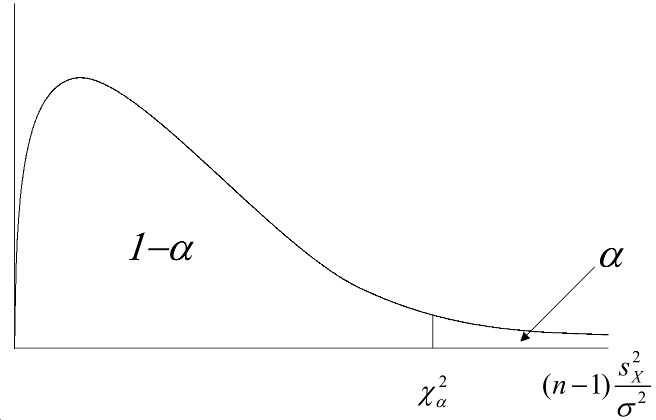
The Other One-sided Confidence Interval for σ^2

$$= \{L, \infty\}$$

$$= \left\{(n-1)\frac{s_X^2}{\chi^2_{\alpha}}, \infty\right\}$$

If we are worried about the lower limit of the confidence interval. How to choose L?

How to choose *L* for One-sided CI



Hypothesis Testing for μ

Prior info. suggests the null hypothesis H_0 or a hypothesized value of μ

$$H_0 : \mu = 10$$

$$H_1 : \mu \neq 10$$

H_0 is true until a strong enough evidence (observed data) suggests that H_0 is unlikely or, alternatively, H_1 is true.

H_1 is also referred to as the Alternative hypothesis

Two-sided Hypo. Testing for μ (1)

- $\overline{X}$ is an estimator(RV) of μ
- Its observed or calculated value (estimate) could be different from the hypothesized value of μ even though H_0 is true.
- If its observed value is not too far from the hypothesized value, then, we can accept that H_0 is true.

Two-sided Hypo. Testing for μ (2)

How far from the hypothesized value is $\overline{X}$ before H_0 will be rejected?

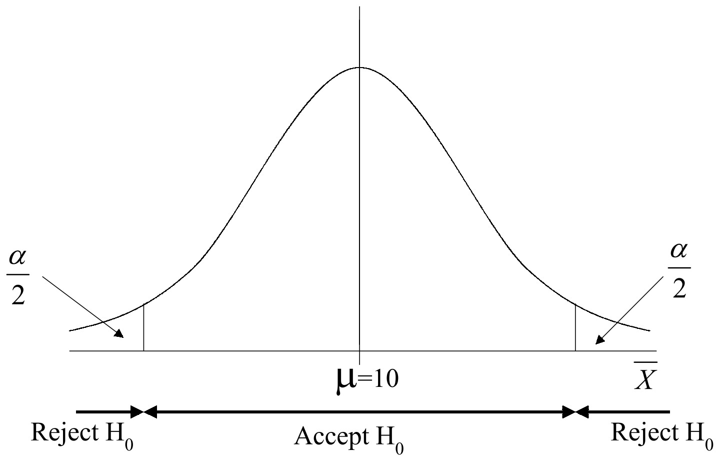
Criterion

Set the probability of Type I Error equal to the Significant Level (α)

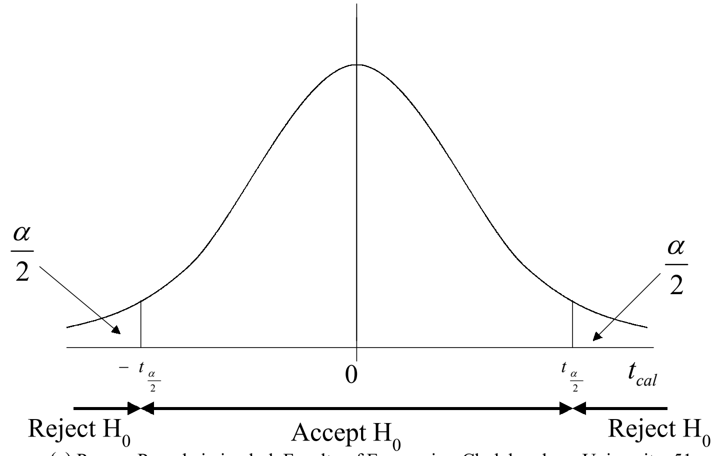
Type I Error

- Type I Error is the event that H_0 has been rejected while it is, in fact, true.
- Low α suggests that the evidence must be very strong before we decide to reject H_0.
- Remember?
  - significant level is chosen by the tester.
  - standard level of significance

Two-sided Hypo. Testing for μ (3)



Two-sided Hypo. Testing for μ (4)



Two-sided Hypo. Testing for μ (5)

Calculate $t_{cal} = \dfrac{\overline{X} - 10}{s_x / \sqrt{n}}$

Criterion compare t_{cal} with t-value from t-table

$$|t_{cal}| \le t_{\frac{\alpha}{2}} ==> accept \;\; H_0$$

$$|t_{cal}| > t_{\frac{\alpha}{2}} ==> reject \;\; H_0$$

One-sided Hypo. Testing for μ (1)

Also called One-tailed

Prior info. suggests that no need to be worried about the other tail(say, $\mu < 10$).

<u>Example</u>

H_0: $\mu = 10$

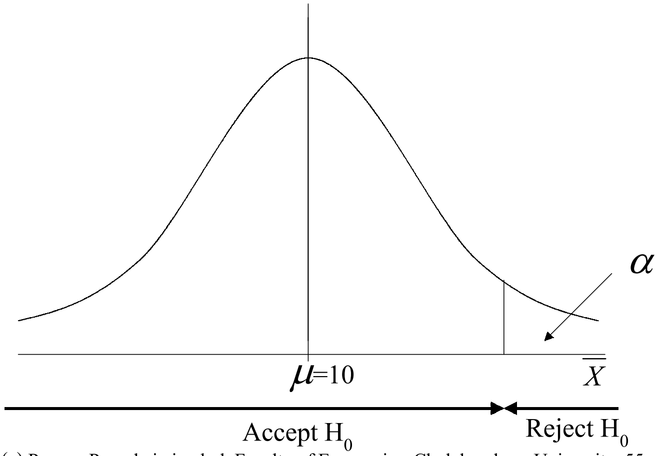
H_1: $\mu > 10$

Do a right-tailed test

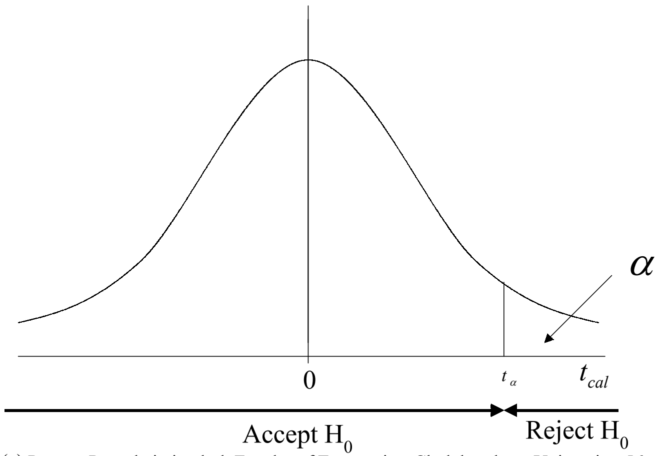
One-sided Hypo. Testing for μ (2)

- Even if the observed/calculated value is less than or equal to the hypothesized value of μ, there is no good reason to choose $\mu > 10$ over $\mu = 10$
- Only if the observed value is larger than the hypothesized value by a large enough margin, then, we can reject H_0.

One-sided Hypo. Testing for μ (3)



One-sided Hypo. Testing for μ (4)



One-sided Hypo. Testing for μ (5)

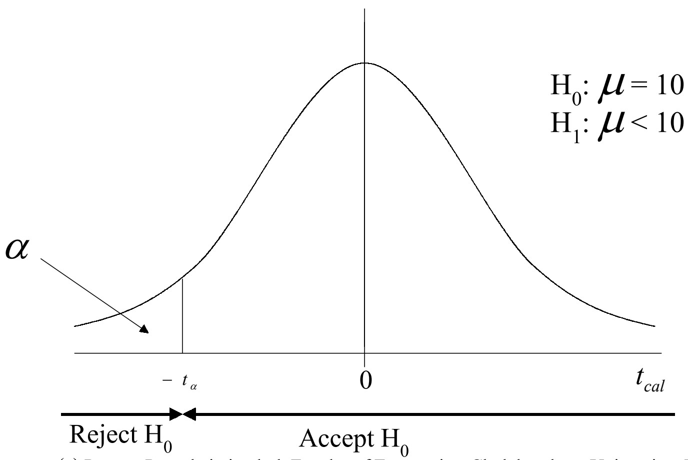
Calculate $t_{cal} = \dfrac{\overline{X} - 10}{s_x / \sqrt{n}}$

Criterion compare t_{cal} with t-value from t-table

$t_{cal} \le t_{\alpha}$ ==> accept H_0

$t_{cal} > t_{\alpha}$ ==> reject H_0

Left-sided Hypo. Testing for μ (1)



Left-sided Hypo. Testing for μ (2)

Calculate $t_{cal} = \dfrac{\overline{X} - 10}{s_x / \sqrt{n}}$

Criterion compare t_{cal} with t-value from t-table

$t_{cal} \ge -t_{\alpha}$ ==> accept H_0

$t_{cal} < -t_{\alpha}$ ==> reject H_0

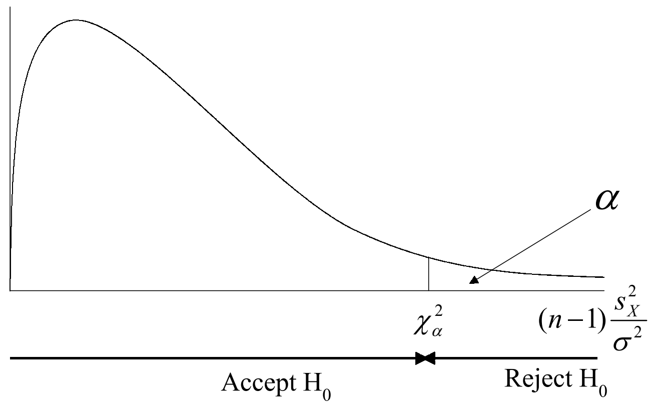
Hypothesis Testing for σ^2 (1)

Variance is always non-negative.
Right-sided test is most interesting since we are more concerned about too high variance, i.e,

H_0: $\sigma^2 = 16$

H_1: $\sigma^2 > 16$

Hypothesis Testing for σ² (2)



Hypothesis Testing for σ² (3)

Calculate $\chi^2_{Cal} = (n-1)\frac{s_X^2}{16}$

Criterion compare χ^2_{Cal} with χ^2-value from χ^2-table

$$\chi^2_{cal} \le \chi^2_\alpha \implies accept\ \ H_0$$

$$\chi^2_{cal} > \chi^2_\alpha \implies reject\ \ H_0$$