Data Analysis

Sergey Vladimirovich Petropavlovsky

National Research University Higher School of Economics
Master's Program "Big Data Systems"

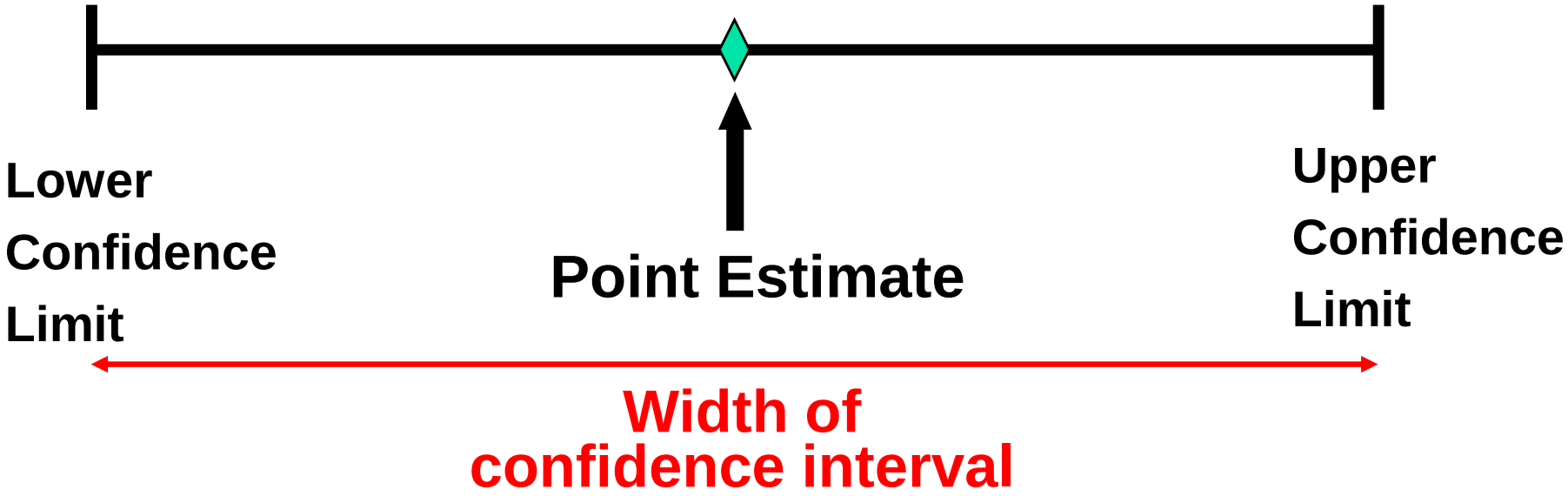
Fall 2018

Plan of Presentation

- Confidence intervals
- Hypothesis testing

Point and Interval Estimates

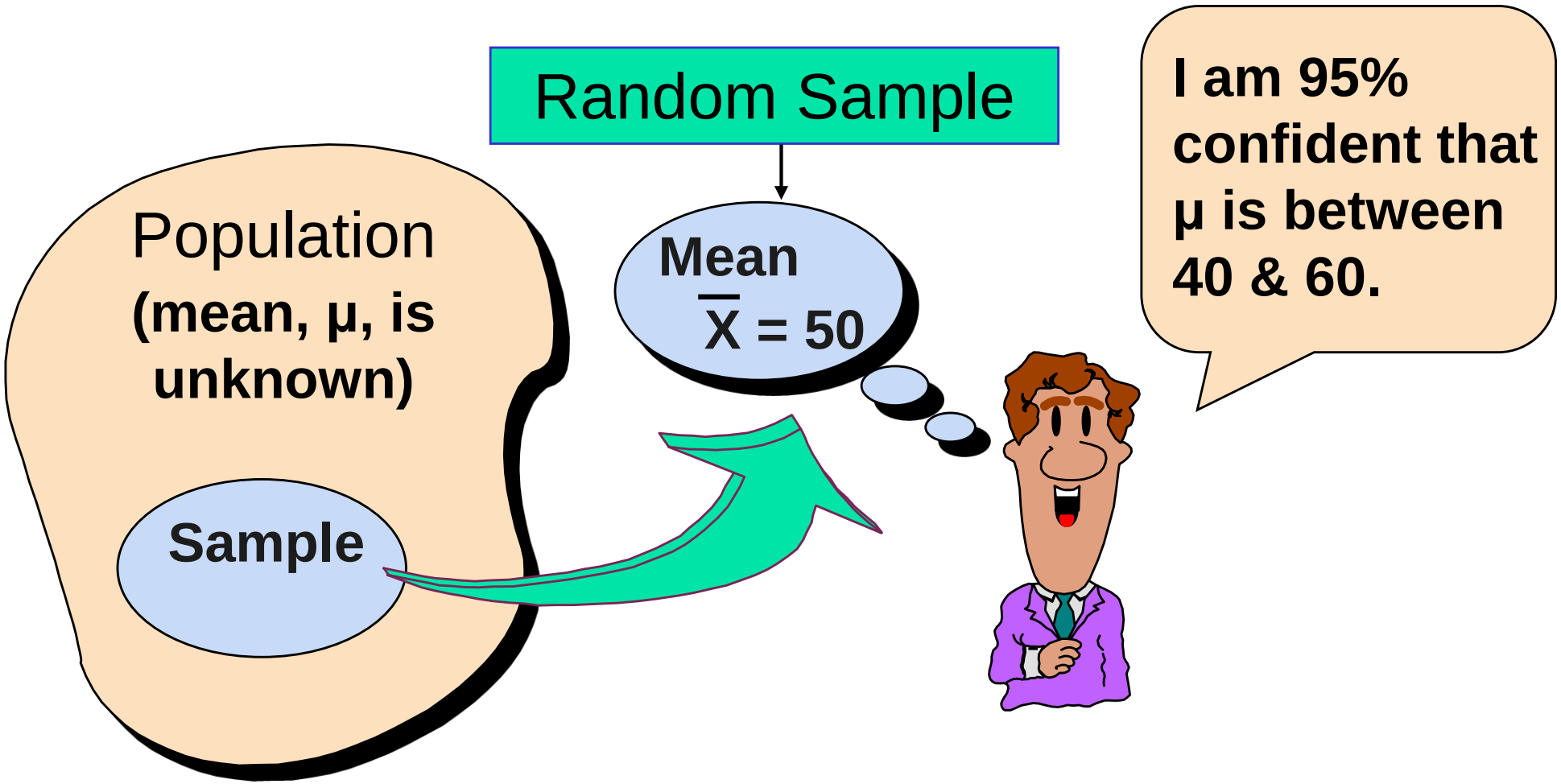
- A point estimate is a single number,
- a confidence interval provides additional information about variability



Confidence Interval for θ and Confidence Level

- If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called a $100(1 - \alpha)\%$ confidence interval of θ.
- The quantity $(1 - \alpha)$ is called the confidence level of the interval (α between 0 and 1)
  - In repeated samples of the population, the true value of the parameter θ would be contained in $100(1 - \alpha)\%$ of intervals calculated this way.
  - The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1 - \alpha)\%$ confidence

Estimation Process



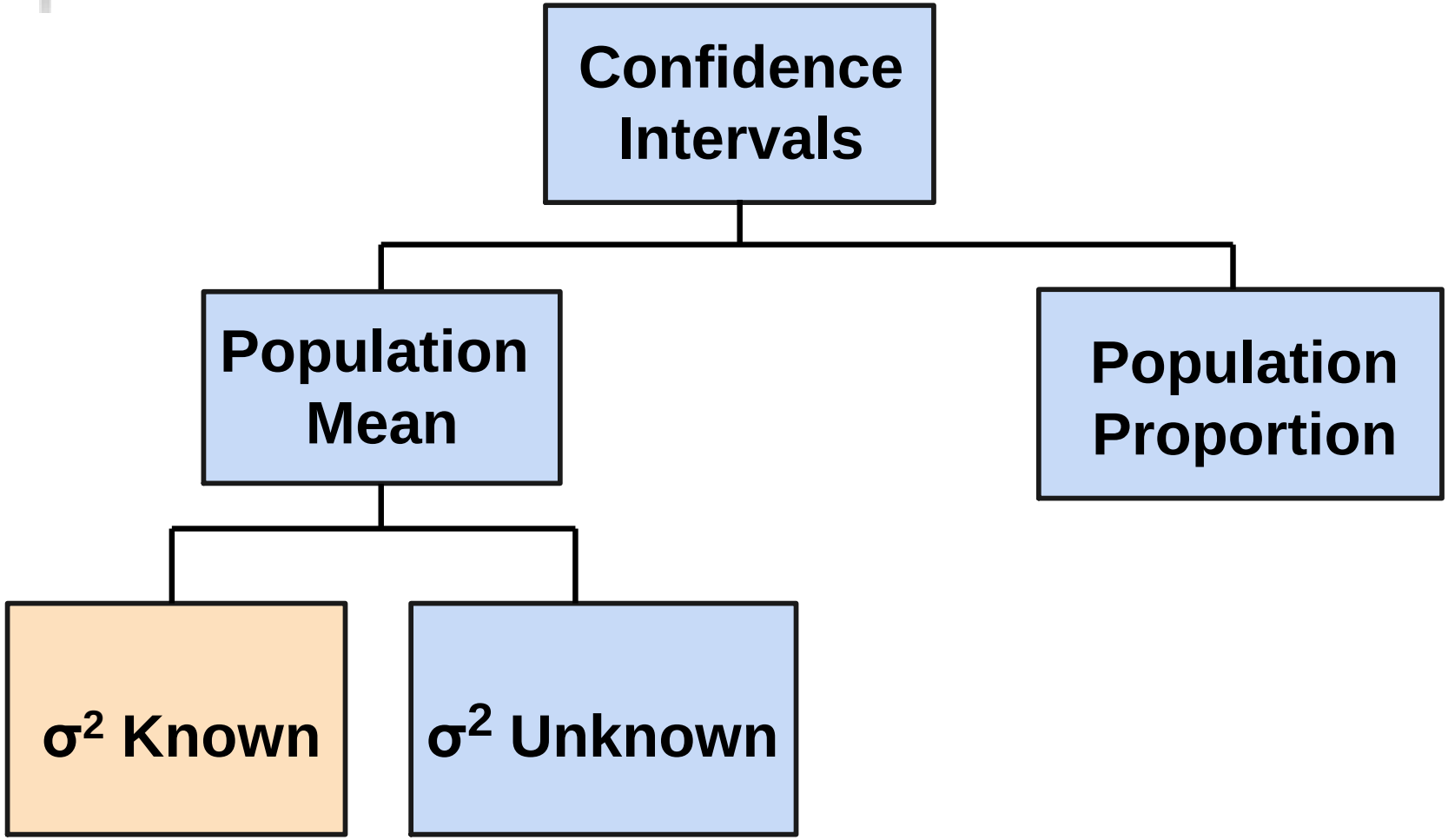
General Formula

- The general formula for all confidence intervals is:

Point Estimate ± (Reliability Factor)(Standard Error)

- The value of the reliability factor depends on the desired level of confidence

Confidence intervals for a single statistic



Confidence Interval for mean (σ^2 known)

- Assumptions
  - Population variance σ^2 is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate:

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

(where $z_{\alpha/2}$ is the normal distribution value for a probability of $\alpha/2$ in each tail)

Reducing the Margin of Error

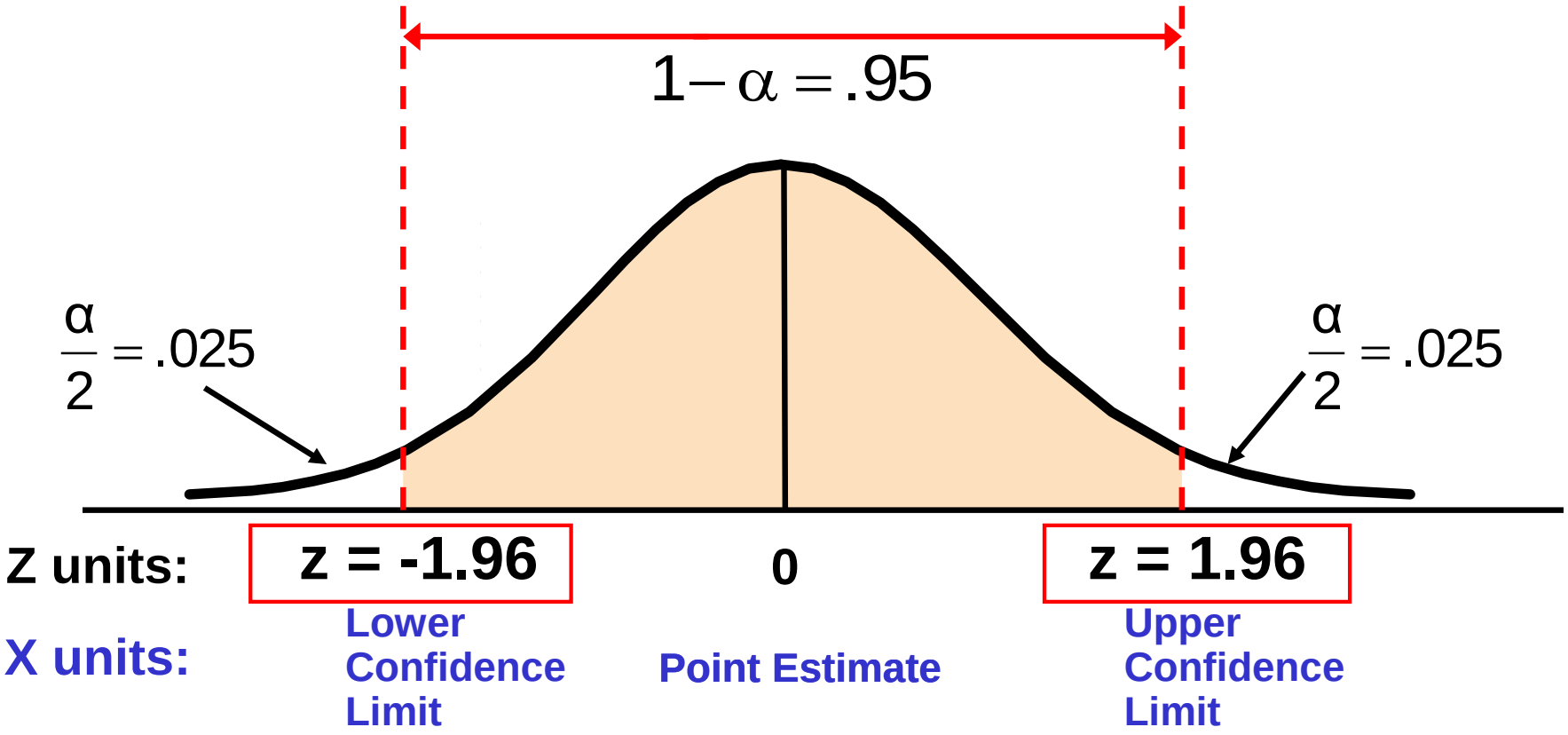
$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced ($\sigma\downarrow$)
- The sample size is increased ($n\uparrow$)
- The confidence level is decreased, $(1 - \alpha) \downarrow$

Finding the Reliability Factor, $z_{\alpha/2}$

- Consider a 95% confidence interval:



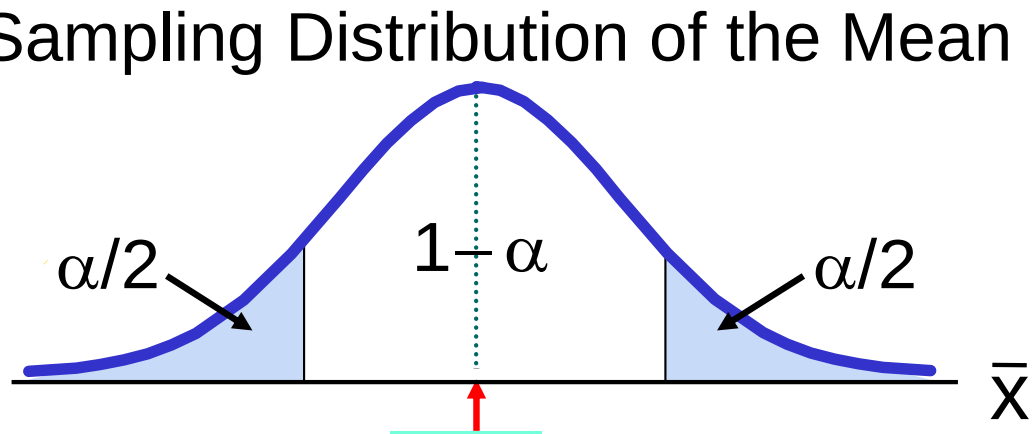
- Find $z_{.025} = \pm 1.96$ from the standard normal distribution table

Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*Confidence Coefficient,* $1-\alpha$	$Z_{\alpha/2}$ *value*
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

Intervals and Level of Confidence

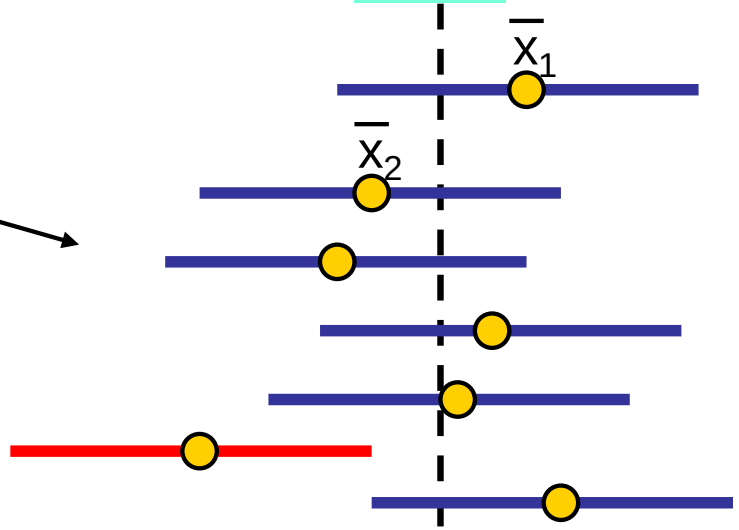


Intervals extend from

$$\bar{x} - z\frac{\sigma}{\sqrt{n}}$$

to

$$\bar{x} + z\frac{\sigma}{\sqrt{n}}$$



Confidence Intervals

100(1-α)% of intervals constructed contain μ;

100(α)% do not.

Example

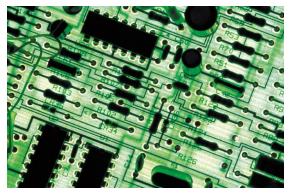
- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is .35 ohms.
- Solution:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\ (.35/\sqrt{11})$$

$$= 2.20 \pm .2068$$

$$1.9932 < \mu < 2.4068$$



Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

Student's t Distribution

- Consider a random sample of n observations
  - with mean $\bar{x}$ and standard deviation s
  - from a normally distributed population with mean μ

- Then the variable

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows the Student's t distribution with (n - 1) degrees of freedom

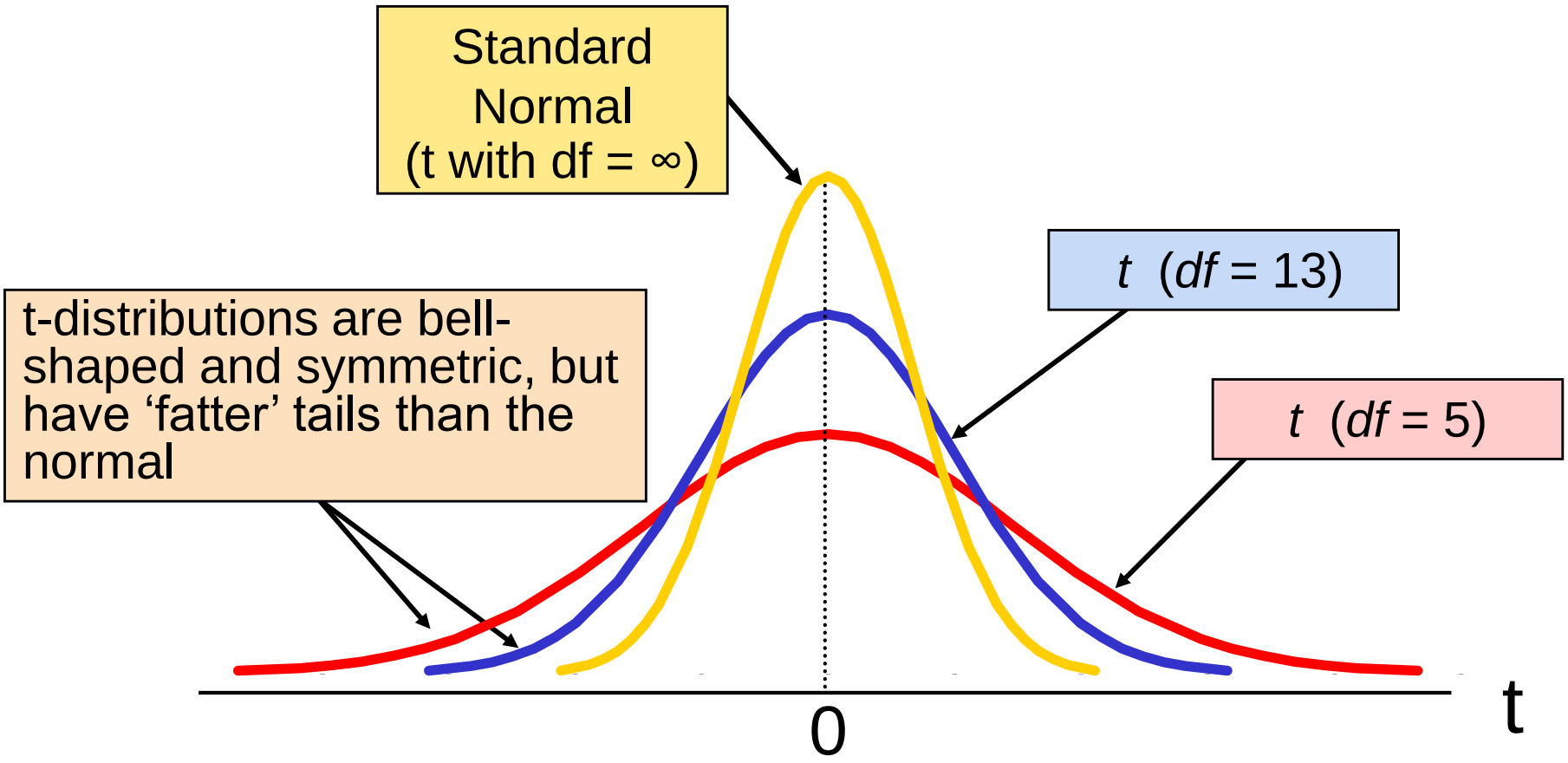
Student's t Distribution (2)

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
- Number of observations that are free to vary after sample mean has been calculated

$$\text{df} = n - 1$$

Student's t Distribution (3)

Note: $t \longrightarrow Z$ as n increases



Confidence Interval for mean (σ^2 unknown)

- If the population standard deviation σ is unknown, we can substitute it with the sample standard deviation, s
- This introduces extra uncertainty since s is variable from sample to sample
- So, we use the t-distribution instead of the normal distribution

Confidence Interval for mean (σ^2 unknown) (2)

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\bar{x} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the critical value of the t distribution with n-1 d.f. and an area of $\alpha/2$ in each tail

$$P(t_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$$

Example

A random sample of n = 25 has $\bar{x} = 50$ and s = 8. Form a 95% confidence interval for μ

- d.f. = n – 1 = 24, so $t_{n-1,\alpha/2} = t_{24,.025} = 2.0639$

The confidence interval is

$$\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

$$50 - (2.0639) \frac{8}{\sqrt{25}} < \mu < 50 + (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$

Confidence Intervals for Mean in R

```
> ozs = c(1.95, 1.80, 2.10, 1.82, 1.75, 2.01, 1.83, 1.90)
> t.test(ozs,conf.level=0.80)
One Sample t-test
data:  ozs
t = 45.25, df = 7, p-value = 6.724e-10
alternative hypothesis:  true mean is not equal to a
80 percent confidence interval:
1.836 1.954
sample estimates:
mean of x
1.895
```

Confidence Intervals for the Population Proportion, P

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- where
  - $z_{\alpha/2}$ is the standard normal value for the level of confidence desired
  - $\hat{p}$ is the sample proportion
  - n is the sample size

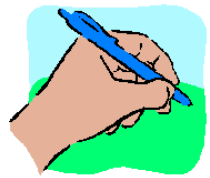
Example

- A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\frac{25}{100} - 1.96\sqrt{\frac{.25(.75)}{100}} < P < \frac{25}{100} + 1.96\sqrt{\frac{.25(.75)}{100}}$$

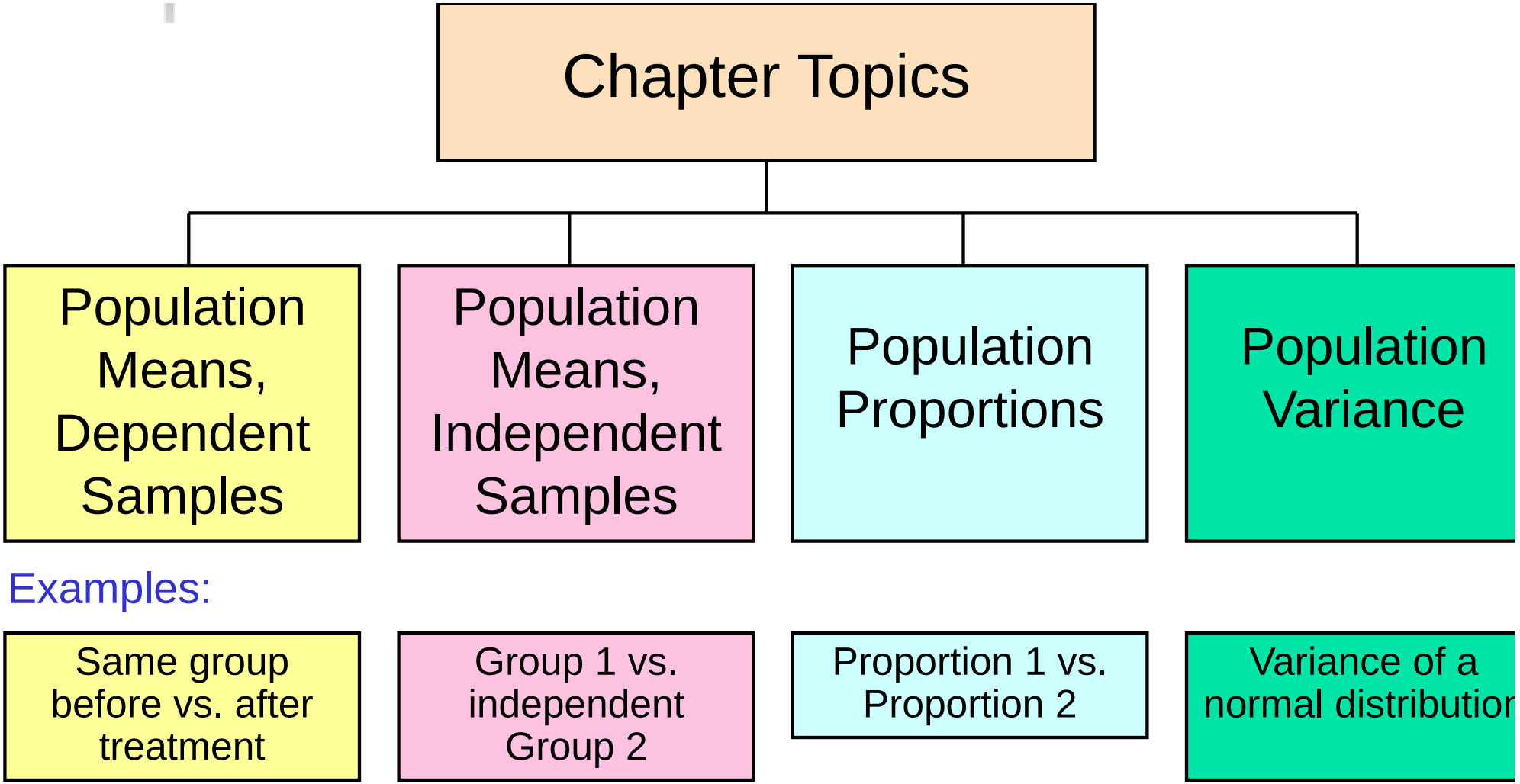
$$0.1651 < P < 0.3349$$



Confidence Intervals for Proportion in R

```
prop.test(466,1013,conf.level=0.95)
1-sample proportions test with continuity correction
data:  466 out of 1013, null probability 0.5
X-squared = 6.318, df = 1, p-value = 0.01195
alternative hypothesis:  true p is not equal to 0.5
95 percent confidence interval:
0.4290 0.4913
sample estimates:
P
0.46
```

Confidence intervals for differences



Dependent Samples

Dependent samples

Tests Means of 2 Related Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use difference between paired values:

$$d_i = x_i - y_i$$

- Eliminates Variation Among Subjects
- Assumptions:
  - Both Populations Are Normally Distributed

Mean Difference

Dependent samples

The i^{th} paired difference is d_i, where

$$d_i = x_i - y_i$$

The point estimate for the population mean paired difference is $\bar{d}$:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

The sample standard deviation is:

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

n is the number of matched pairs in the sample

Confidence Interval for Mean Difference

Dependent samples

The confidence interval for difference between population means, μ_d , is

$$\bar{d} - t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

Where

n = the sample size

(number of matched pairs in the paired sample)

Confidence Interval for Mean Difference (2)

Dependent samples

- The margin of error is

$$ME = t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

- $t_{n-1,\alpha/2}$ is the value from the Student's t distribution with (n – 1) degrees of freedom for which

$$P(t_{n-1} > t_{n-1,\alpha/2}) = \frac{\alpha}{2}$$

Paired Samples Example

- Six people sign up for a weight loss program. You collect the following data:

Person	Weight: Before (x)	After (y)	Difference, d_i
1	136	125	11
2	205	195	10
3	157	150	7
4	138	140	- 2
5	175	165	10
6	166	160	6
			42

$$\bar{d} = \frac{\sum d_i}{n}$$

$$= 7.0$$

$$S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

$$= 4.82$$

Paired Samples Example (2)

- For a 95% confidence level, the appropriate t value is $t_{n-1,\alpha/2} = t_{5,.025} = 2.571$
- The 95% confidence interval for the difference between means, μ_d, is

$$\bar{d} - t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}}$$

$$7 - (2.571)\frac{4.82}{\sqrt{6}} < \mu_d < 7 + (2.571)\frac{4.82}{\sqrt{6}}$$

$$-1.94 < \mu_d < 12.06$$

Since this interval contains zero, we cannot be 95% confident, given this limited data, that the weight loss program helps people lose weight

Difference Between Two Means

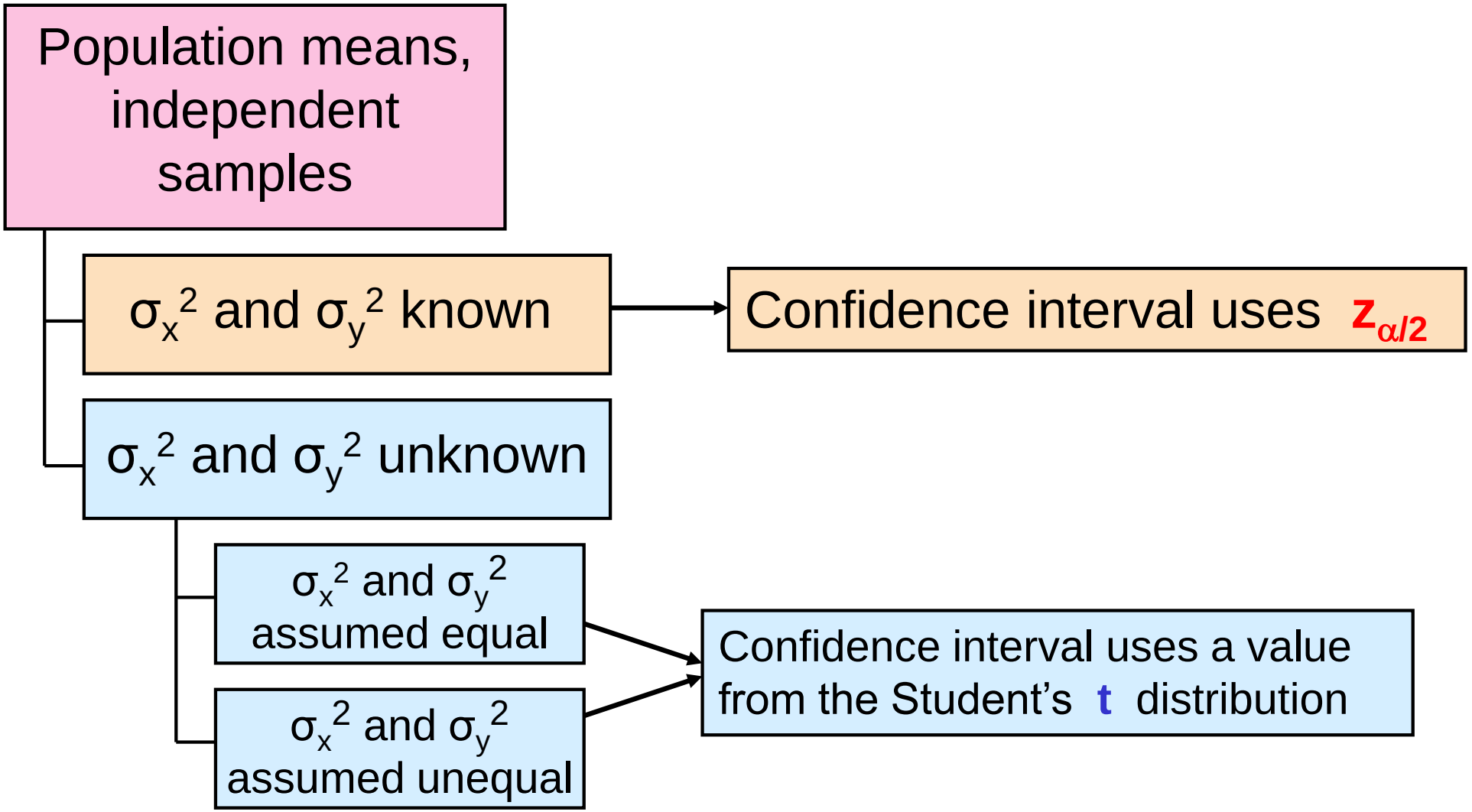
Population means, independent samples

Goal: Form a confidence interval for the difference between two population means, $\mu_x - \mu_y$

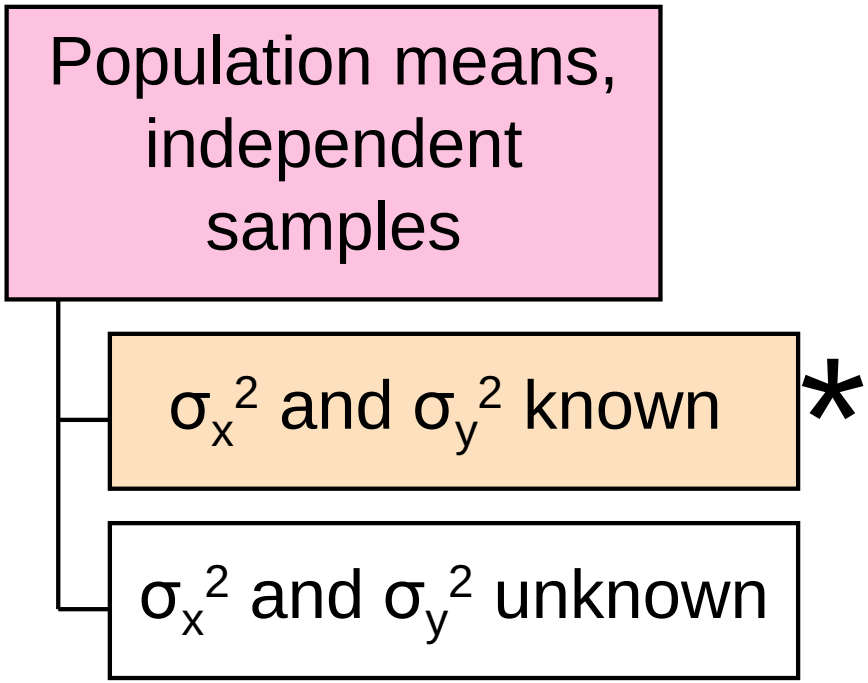
- Different data sources
  - Unrelated
  - Independent
    - Sample selected from one population has no effect on the sample selected from the other population
- The point estimate is the difference between the two sample means:

$$\overline{x} - \overline{y}$$

Difference Between Two Means (2)



σ_x^2 and σ_y^2 known



Assumptions:

- Samples are randomly and independently drawn
- both population distributions are normal
- Population variances are known

σ_x^2 and σ_y^2 known (2)

- Population means, independent samples
  - σ_x^2 and σ_y^2 known *
  - σ_x^2 and σ_y^2 unknown

When σ_x and σ_y are known and both populations are normal, the variance of $\overline{X} - \overline{Y}$ is

$$\sigma^2_{\overline{X}-\overline{Y}} = \frac{\sigma_x^{\,2}}{n_x} + \frac{\sigma_y^{\,2}}{n_y}$$

…and the random variable

$$Z = \frac{(\overline{x}-\overline{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_x^2}{n_X}+\frac{\sigma_y^2}{n_Y}}}$$

has a standard normal distribution

Confidence interval for the difference: σ_x^2 and σ_y^2 known

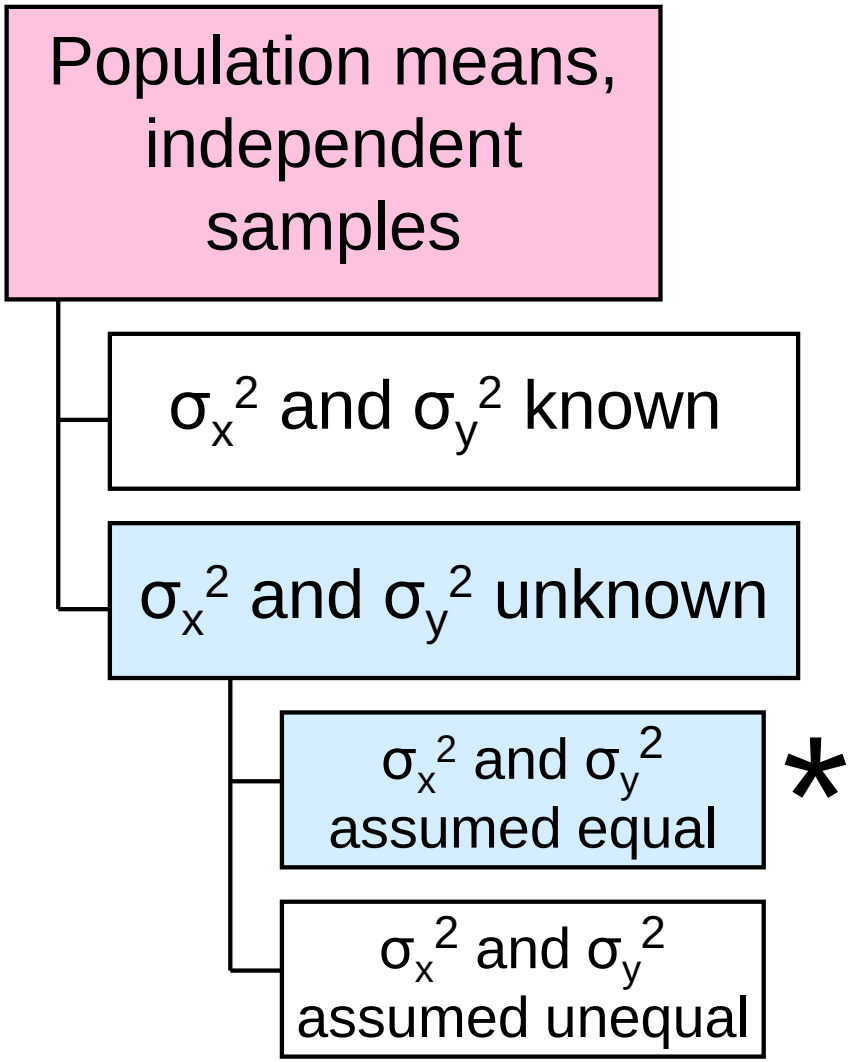
Population means, independent samples

- σ_x^2 and σ_y^2 known
- σ_x^2 and σ_y^2 unknown

* The confidence interval for $\mu_x - \mu_y$ is:

$$(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_x}+\frac{\sigma_Y^2}{n_y}} < \mu_X - \mu_Y < (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_x}+\frac{\sigma_Y^2}{n_y}}$$

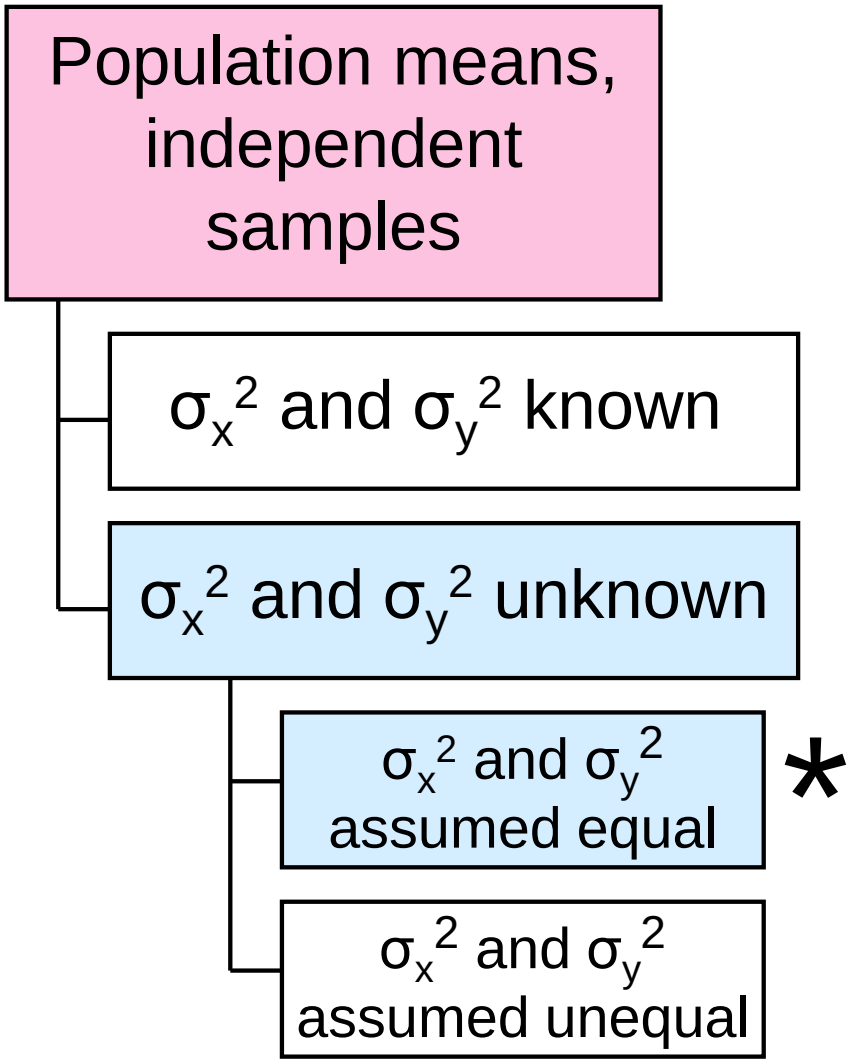
σ_x^2 and σ_y^2 unknown, assumed equal



Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown but assumed equal

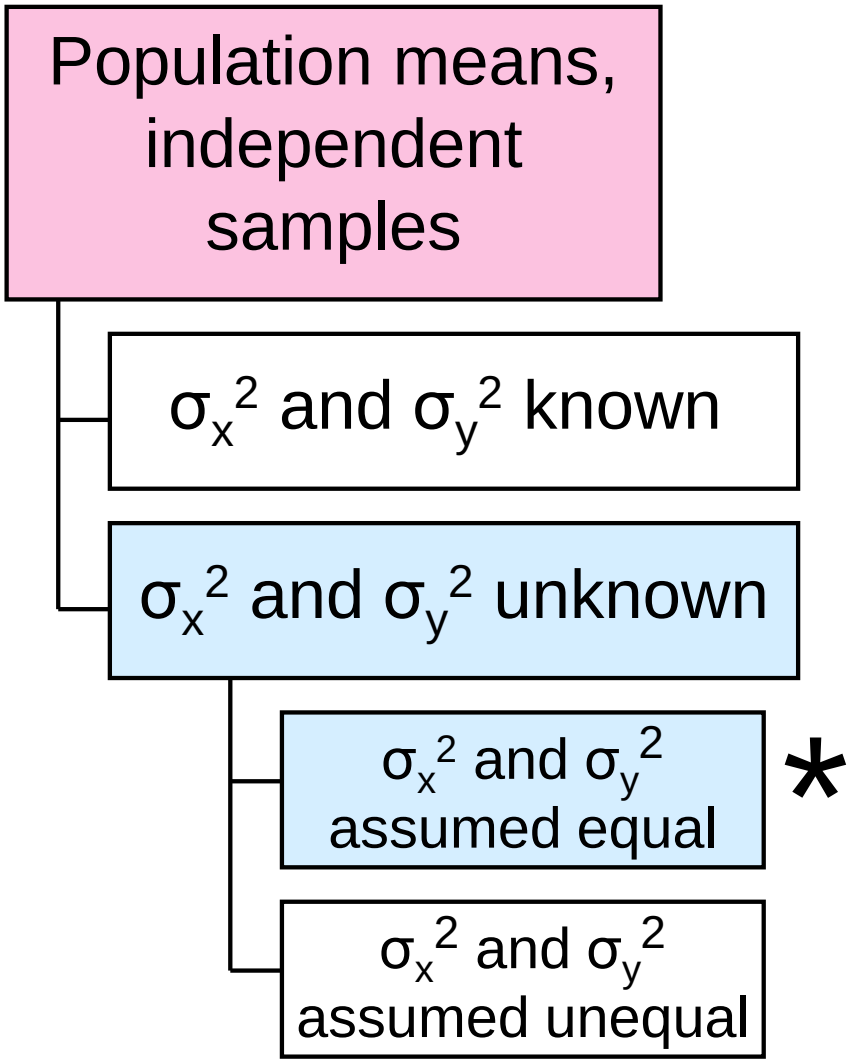
σ_x^2 and σ_y^2 unknown, assumed equal (2)



Forming interval estimates:

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ
- use a t value with $(n_x + n_y - 2)$ degrees of freedom

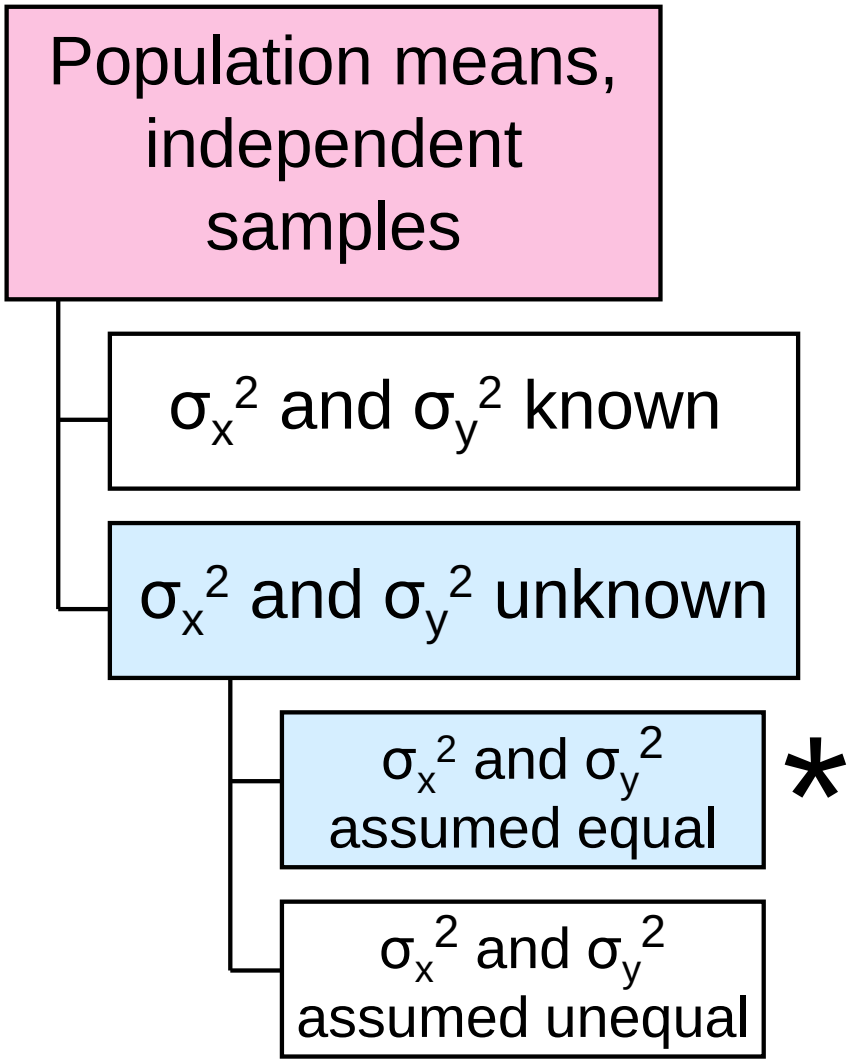
σ_x^2 and σ_y^2 unknown, assumed equal (3)



The pooled variance is

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

Confidence interval, σ_x^2 and σ_y^2 unknown, assumed equal



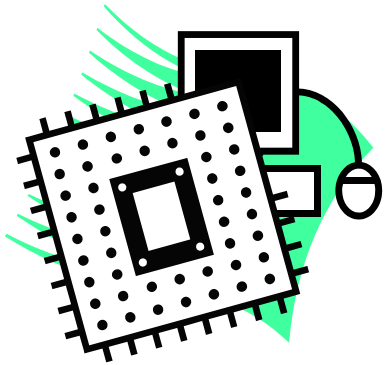
Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown but assumed equal

Pooled Variance Example

You are testing two computer processors for speed. Form a confidence interval for the difference in CPU speed. You collect the following speed data (in Mhz):

	CPU_x	CPU_y
Number Tested	**17**	**14**
Sample mean	**3004**	**2538**
Sample std dev	**74**	**56**



Assume both populations are normal with equal variances, and use 95% confidence

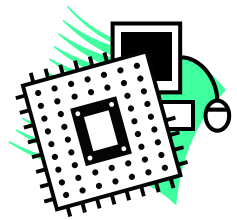
Calculating the Pooled Variance

The pooled variance is:

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{(n_x - 1) + (n_y - 1)} = \frac{(17 - 1)74^2 + (14 - 1)56^2}{(17 - 1) + (14 - 1)} = \boxed{4427.03}$$

The t value for a 95% confidence interval is:

$$\boxed{t_{n_x + n_y - 2, \alpha/2} = t_{29, 0.025} = 2.045}$$



Calculating the Confidence Limits

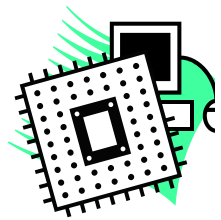
- The 95% confidence interval is

$$(\bar{x}-\bar{y})-t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}<\mu_X-\mu_Y<(\bar{x}-\bar{y})+t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$$

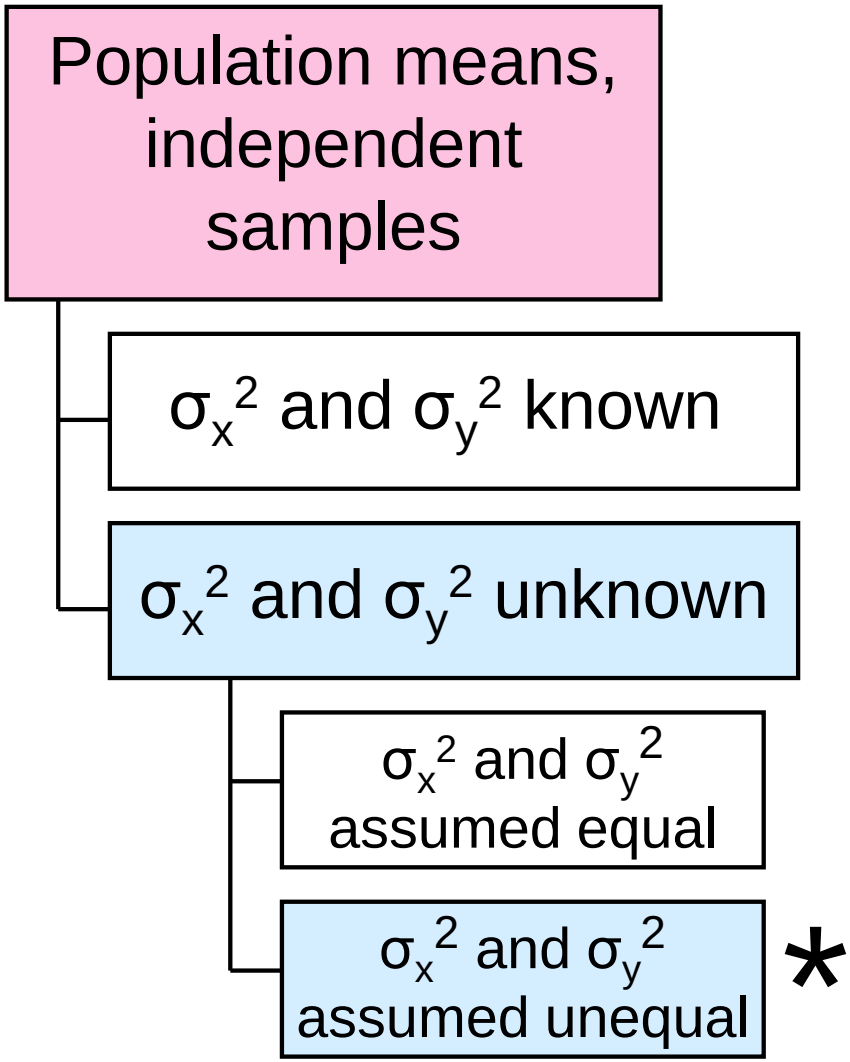
$$(3004-2538)-(2.054)\sqrt{\frac{4427.03}{17}+\frac{4427.03}{14}}<\mu_X-\mu_Y<(3004-2538)+(2.054)\sqrt{\frac{4427.03}{17}+\frac{4427.03}{14}}$$

$$\boxed{416.69<\mu_X-\mu_Y<515.31}$$

We are 95% confident that the mean difference in CPU speed is between 416.69 and 515.31 Mhz.



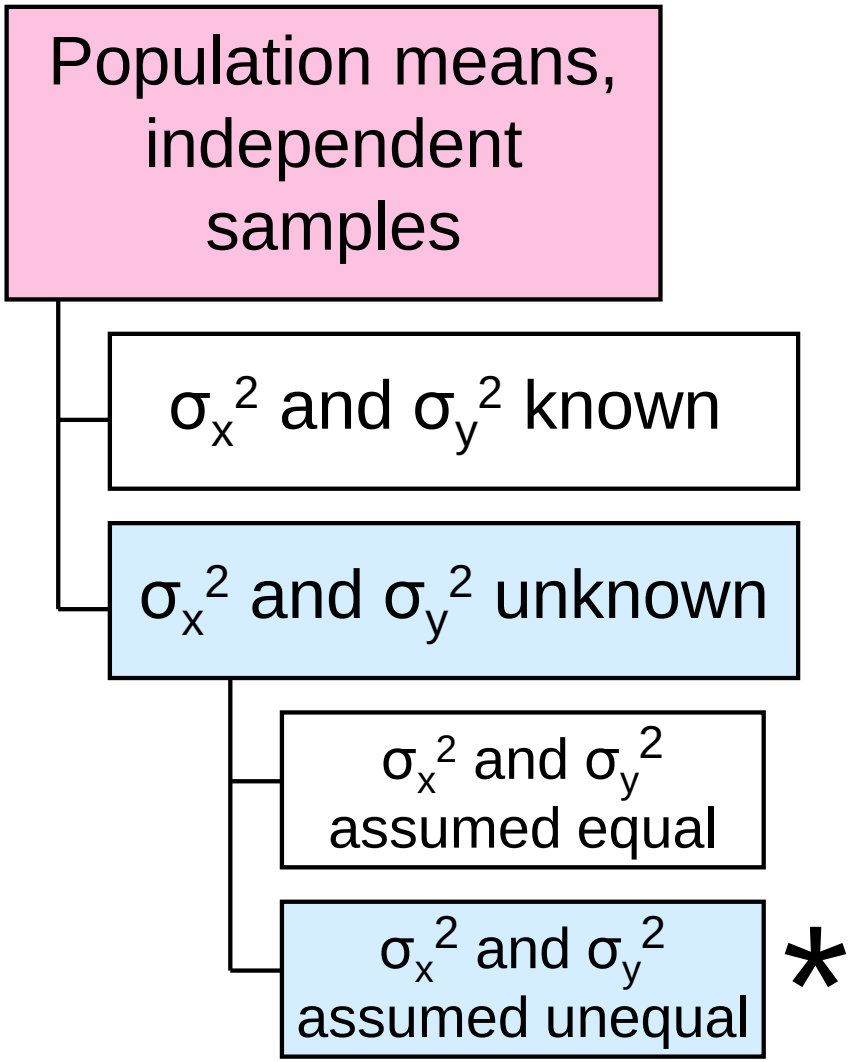
σ_x^2 and σ_y^2 unknown, assumed unequal



Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown and assumed unequal

σ_x^2 and σ_y^2 unknown, assumed unequal (2)



Forming interval estimates:

- The population variances are assumed unequal, so a pooled variance is not appropriate
- use a t value with ν degrees of freedom, where

$$\nu = \frac{\left[\left(\frac{s_x^2}{n_x}\right)+\left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$$

Confidence intervals σ_x^2 and σ_y^2 unknown, assumed unequal

- σ_x^2 and σ_y^2 unknown
  - σ_x^2 and σ_y^2 assumed equal
  - σ_x^2 and σ_y^2 assumed unequal *

The confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x}-\bar{y}) - t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}} < \mu_X - \mu_Y < (\bar{x}-\bar{y}) + t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

Where

$$\nu = \frac{\left[\left(\frac{s_x^2}{n_x}\right)+\left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$$

Two Population Proportions

Population proportions

Goal: Form a confidence interval for the difference between two population proportions, $P_x - P_y$

Assumptions:

Both sample sizes are large (generally at least 40 observations in each sample)

The point estimate for the difference is $\hat{p}_x - \hat{p}_y$

Two Population Proportions (2)

Population proportions

- The random variable

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}$$

is approximately normally distributed

Confidence Interval for Two Population Proportions

Population proportions

The confidence limits for $P_x - P_y$ are:

$$(\hat{p}_x - \hat{p}_y) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

Two Population Proportions: Example

Form a 90% confidence interval for the difference between the proportion of men and the proportion of women who have college degrees.

- In a random sample, 26 of 50 men and 28 of 40 women had an earned college degree

Two Population Proportions: Example (2)

Men: $\hat{p}_x = \frac{26}{50} = 0.52$

Women: $\hat{p}_y = \frac{28}{40} = 0.70$

$$\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = \sqrt{\frac{0.52(0.48)}{50} + \frac{0.70(0.30)}{40}} = 0.1012$$

For 90% confidence, $Z_{\alpha/2} = 1.645$

Two Population Proportions: Example (3)

The confidence limits are:

$$(\hat{p}_x - \hat{p}_y) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

$$= (.52 - .70) \pm 1.645\,(0.1012)$$

so the confidence interval is

$$-0.3465 < P_x - P_y < -0.0135$$

Since this interval does not contain zero we are 90% confident that the two proportions are not equal

Confidence Intervals for Difference Between Two Population Proportions in R

```
> prop.test(x=c(560,570), n=c(1000,1200), conf.level=0.95)
2-sample test for equality of proportions with continuity
correction
data:  c(560, 570) out of c(1000, 1200)
X-squared=15.44, df=1, p-value=8.53e-05
alternative hypothesis:
two sided 95 percent confidence interval:
0.04231 0.12769
sample estimates:
prop 1 prop 2
0.560 0.475
```

Confidence Intervals for the Population Variance

Population Variance

- Goal: Form a confidence interval for the population variance, σ^2
- The confidence interval is based on the sample variance, s^2
- Assumed: the population is normally distributed

Confidence Intervals for the Population Variance (2)

Population Variance

The random variable

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with (n – 1) degrees of freedom

The chi-square value $\chi^2_{n-1,\alpha}$ denotes the number for which

$$P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$$

Confidence Intervals for the Population Variance (3)

Population Variance

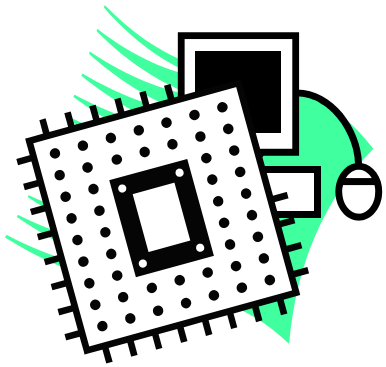
The (1 - α)% confidence interval for the population variance is

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

Confidence Intervals for the Population Variance: Example

You are testing the speed of a computer processor.
You collect the following data (in Mhz):

	CPU_x
Sample size	**17**
Sample mean	**3004**
Sample std dev	**74**



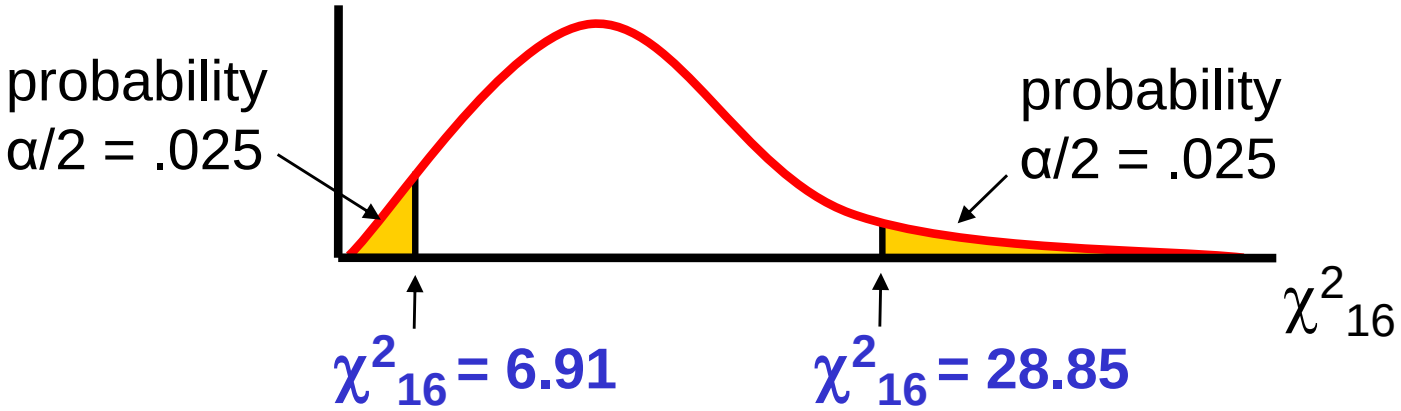
Assume the population is normal.
Determine the 95% confidence interval for σ_x^2

Finding the χ^2 Values

- n = 17 so the chi-square distribution has (n – 1) = 16 degrees of freedom
- α = 0.05, so use the the chi-square values with area 0.025 in each tail:

$$\chi^2_{n-1,\alpha/2} = \chi^2_{16,0.025} = 28.85$$

$$\chi^2_{n-1,1-\alpha/2} = \chi^2_{16,0.975} = 6.91$$



Calculating the Confidence Limits

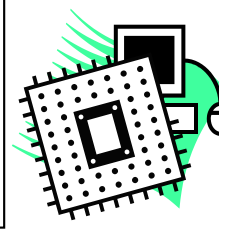
- The 95% confidence interval is

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

$$\frac{(17-1)(74)^2}{28.85} < \sigma^2 < \frac{(17-1)(74)^2}{6.91}$$

$$3037 < \sigma^2 < 12683$$

Converting to standard deviation, we are 95% confident that the population standard deviation of CPU speed is between 55.1 and 112.6 Mhz



Confidence Interval for the Population Variance in R

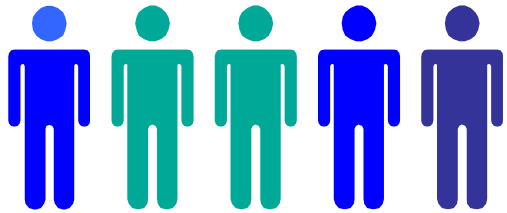
```
> s2 = 12; n = 10
> alpha = .05
> lstar = qchisq(alpha/2, df = n-1)
> rstar = qchisq(1-alpha/2, df = n-1)
> (n-1)*s2 * c(1/rstar,1/lstar) # CI for sigma squared
[1] 5.677 39.994
> sqrt((n-1)*s2 * c(1/rstar,1/lstar)) # CI for sigma
[1] 2.383 6.324
```

Hypothesis Testing Process

Claim: the population mean age is 50. (Null Hypothesis: $H_0: \mu = 50$)

Population

Now select a random sample



Sample

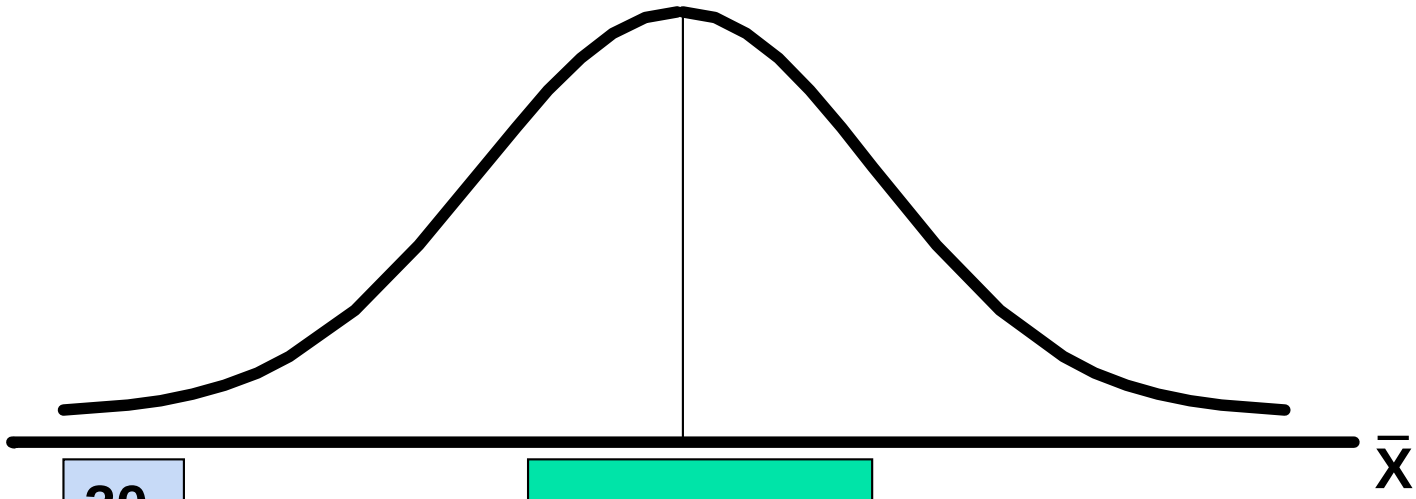
Suppose the sample mean age is 20: $\overline{X} = 20$

Is $\overline{X}=20$ likely if $\mu = 50$?

If not likely,

REJECT Null Hypothesis

Reason for Rejecting H_0



If it is unlikely that we would get a sample mean of this value ...

... if in fact this were the population mean…

... then we reject the null hypothesis that μ = 50.

Level of Significance and the Rejection Region

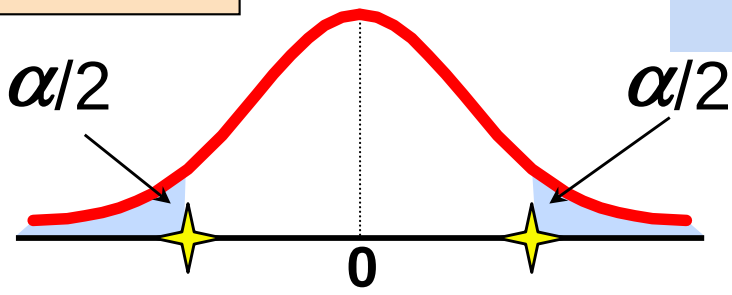
Level of significance = α

✦ **Represents critical value**

H_0: $\mu = 3$

H_1: $\mu \neq 3$

Two-tail test

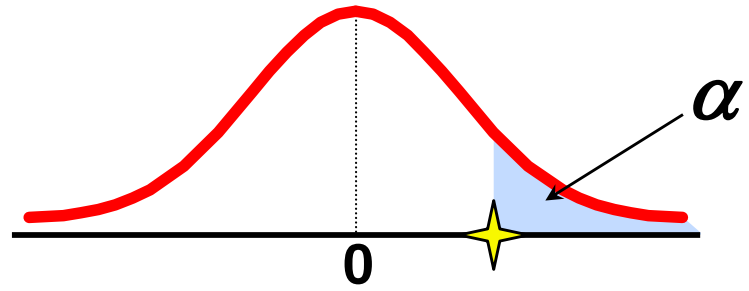


Rejection region is shaded

H_0: $\mu \leq 3$

H_1: $\mu > 3$

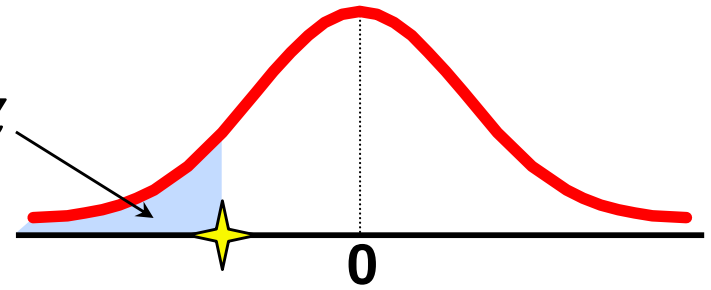
Upper-tail test



H_0: $\mu \geq 3$

H_1: $\mu < 3$

Lower-tail test



Outcomes and Probabilities

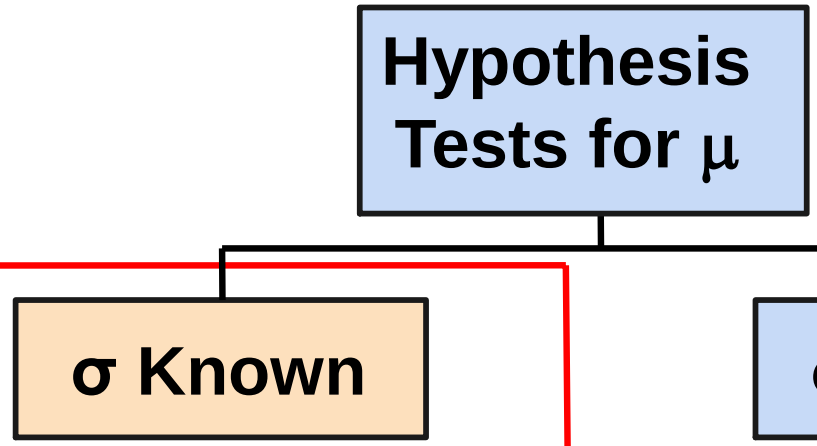
Key:
Outcome
(Probability)

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject **H_0**	**No error** **$(1 - \alpha)$**	**Type II Error** **(β)**
Reject **H_0**	**Type I Error** **(α)**	**No Error** **$(1 - \beta)$**

Test of Hypothesis for the Mean (σ known)

- Convert sample result ($\bar{x}$) to a z value



Consider the test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

(Assume the population is normal)

The decision rule is:

$$\text{Reject } H_0 \text{ if } z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$$

Decision Rule

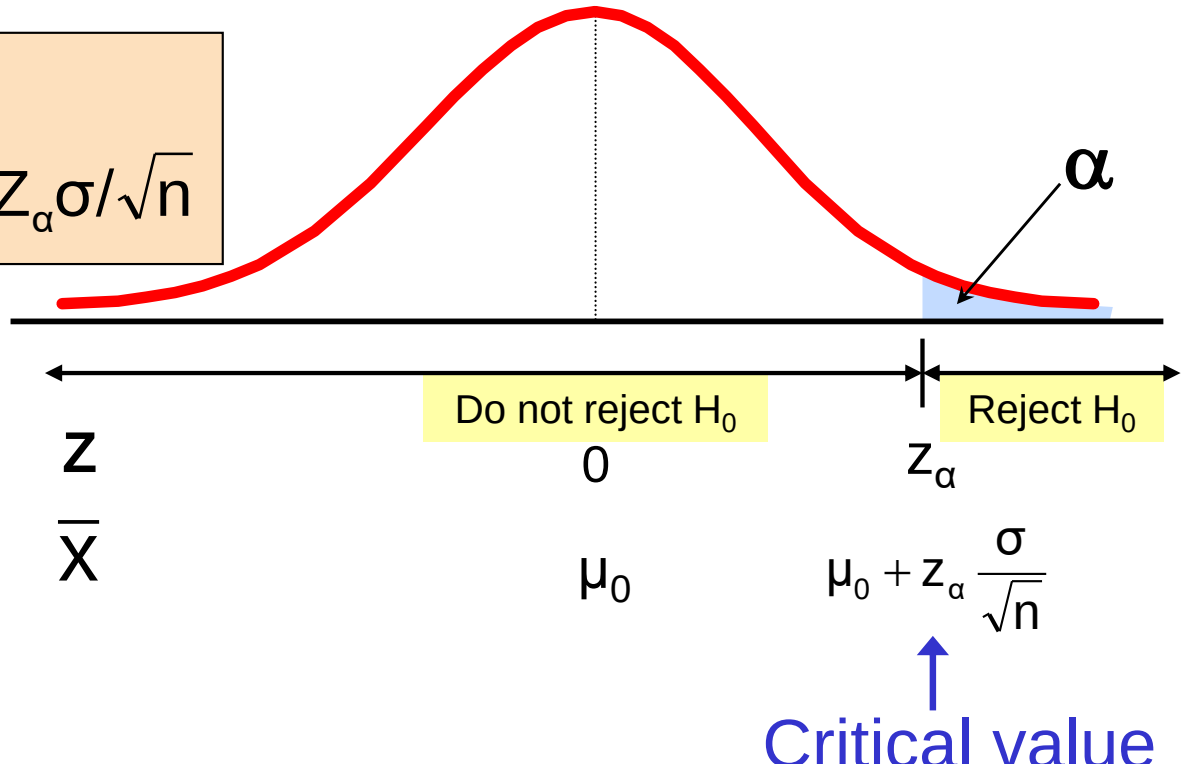
Reject H_0 if $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} > z_\alpha$

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

Alternate rule:

RejectH_0 if $\bar{X} > \mu_0 + Z_\alpha \sigma/\sqrt{n}$



p-Value Approach to Testing

- p-value: Probability of obtaining a test statistic more extreme ($\leq$ or $\geq$) than the observed sample value given H_0 is true
  - Also called observed level of significance
  - Smallest value of α for which H_0 can be rejected

p-Value Approach to Testing (2)

- Convert sample result (e.g., $\bar{x}$) to test statistic (e.g., z statistic)
- Obtain the p-value
  - For an upper tail test:

$$\text{p-value} = P\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \text{, given that } H_0 \text{ is true}\right)$$

$$= P\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0\right)$$

- Decision rule: compare the p-value to α

  - If p-value $<$ α , reject H_0
  - If p-value $\geq$ α , do not reject H_0

Hypothesis Testing Example

Test the claim that the true mean # of TV sets in US homes is equal to 3. (Assume σ = 0.8)

- State the appropriate null and alternative hypotheses
  - H_0: μ = 3 , H_1: μ ≠ 3 (This is a two tailed test)
- Specify the desired level of significance
  - Suppose that α = .05 is chosen for this test
- Choose a sample size
  - Suppose a sample of size n = 100 is selected

Hypothesis Testing Example (2)

- Determine the appropriate technique
  - σ is known so this is a z test
- Set up the critical values
  - For $\alpha = .05$ the critical z values are ± 1.96
- Collect the data and compute the test statistic
  - Suppose the sample results are

    $n = 100$, $\bar{x} = 2.84$ ($\sigma = 0.8$ is assumed known)

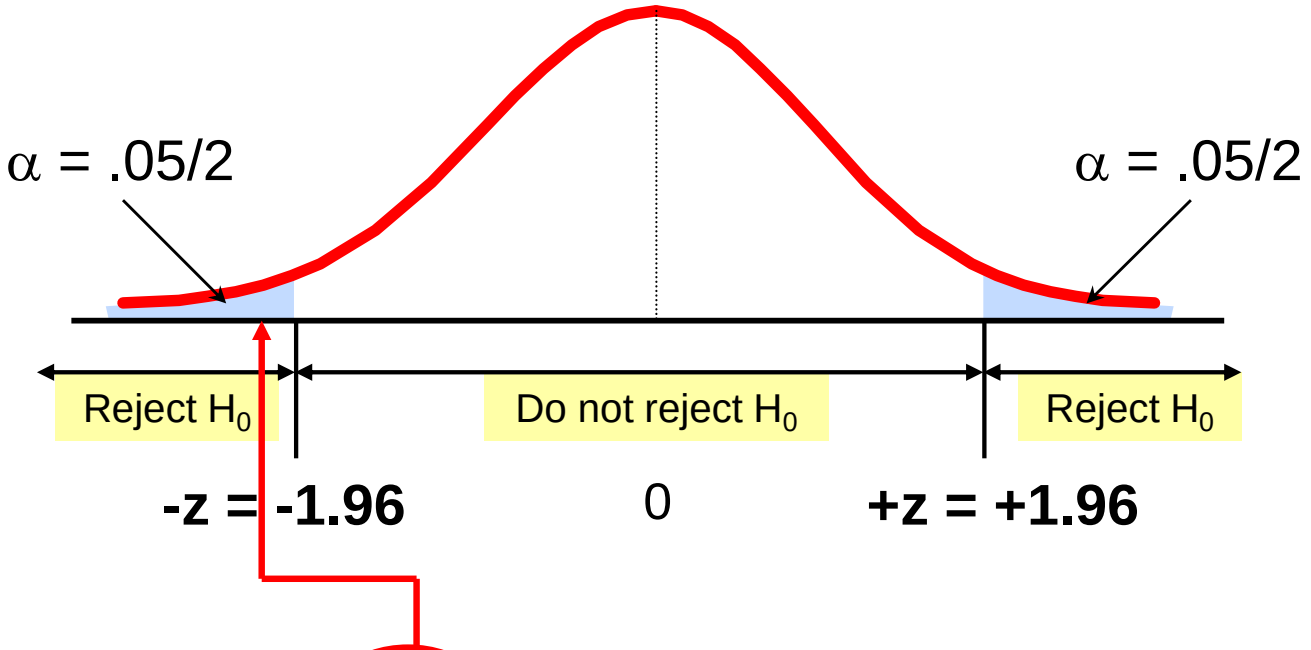
So the test statistic is:

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$

Hypothesis Testing Example (3)

- Is the test statistic in the rejection region?

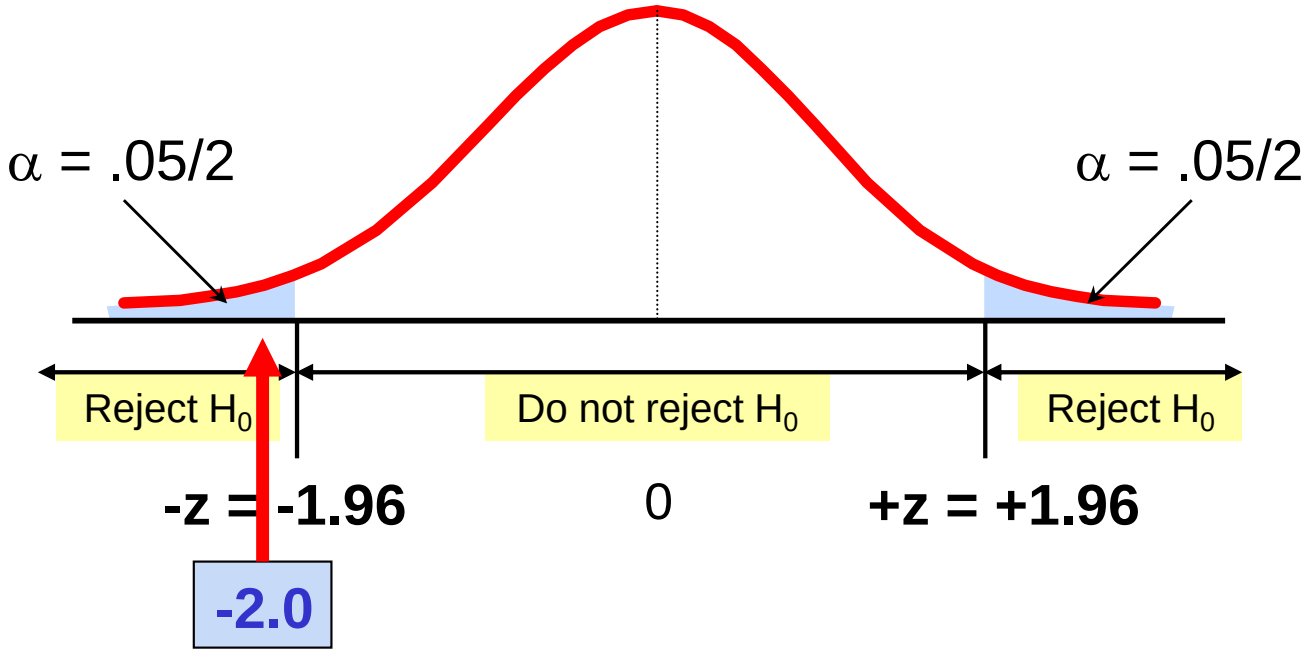
Reject H_0 if $z < -1.96$ or $z > 1.96$; otherwise do not reject H_0



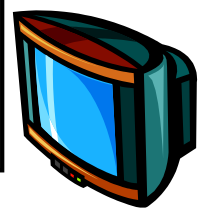
Here, $z = -2.0 < -1.96$, so the test statistic is in the rejection region

Hypothesis Testing Example (4)

- Reach a decision and interpret the result



Since $z = -2.0 < -1.96$, we reject the null hypothesis and conclude that there is sufficient evidence that the mean number of TVs in US homes is not equal to 3



Hypothesis Testing Example: p-value

- Example: How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is $\mu = 3.0$?

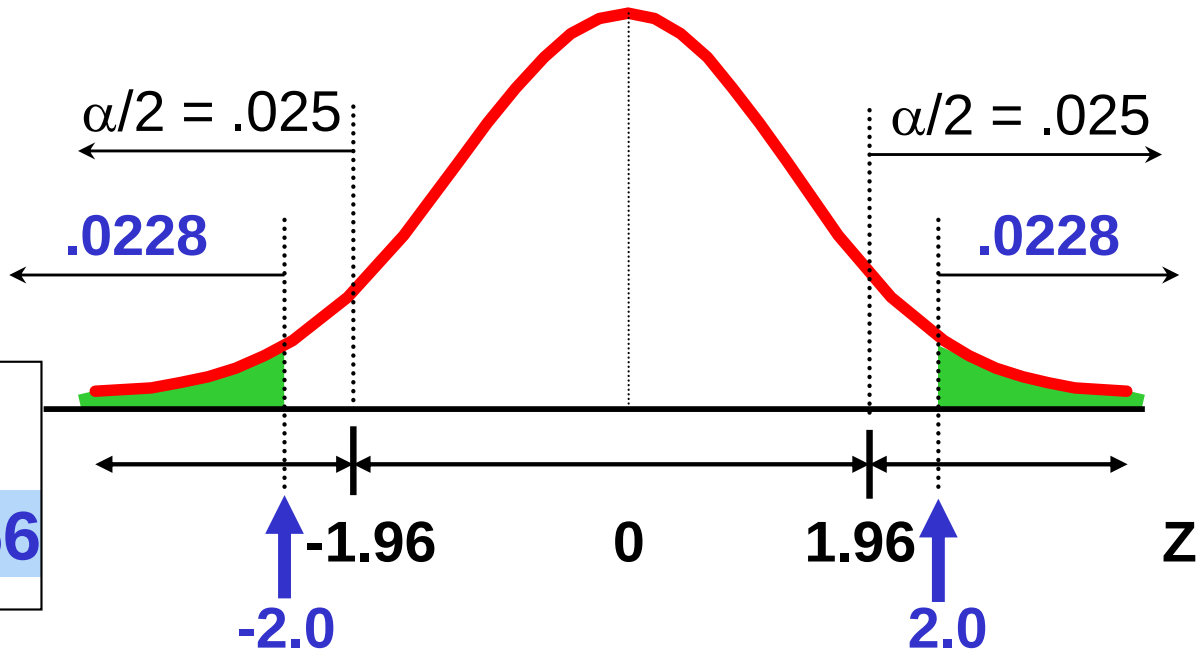
$\overline{x} = 2.84$ is translated to a z score of $z = -2.0$

$P(z < -2.0) = .0228$

$P(z > 2.0) = .0228$

p-value

= .0228 + .0228 = .0456

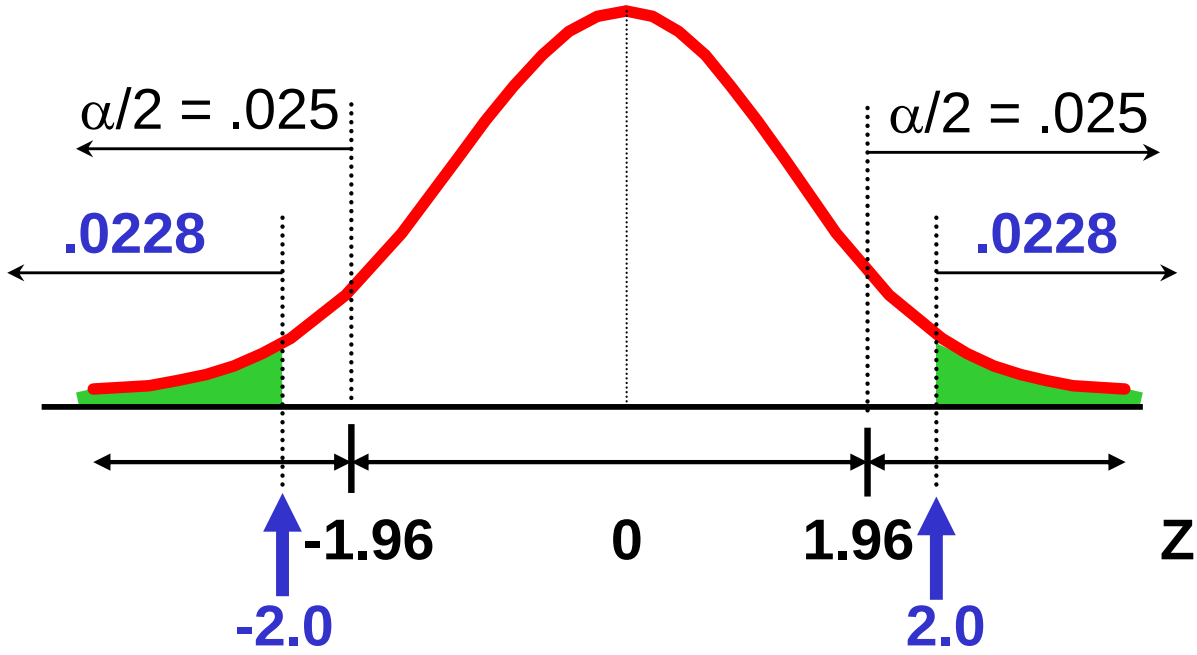


Hypothesis Testing Example: p-value (2)

- Compare the p-value with α
  - If p-value $<$ α , reject H_0
  - If p-value $\geq$ α , do not reject H_0

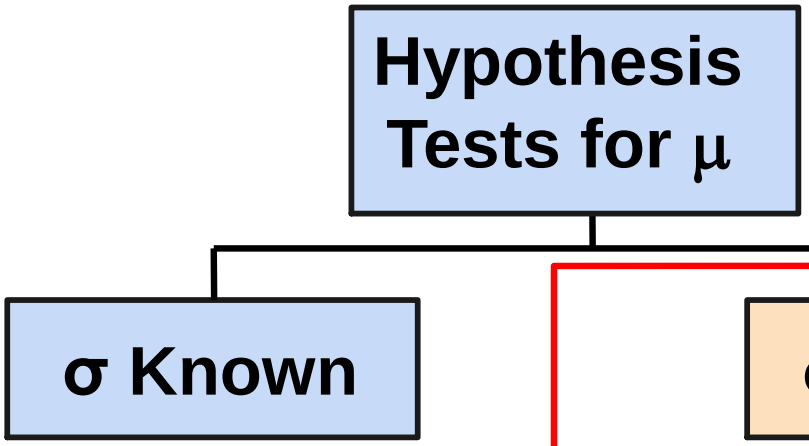
Here: p-value = .0456
α = .05

Since .0456 < .05, we reject the null hypothesis



t-Test of Hypothesis for the Mean (σ unknown)

- Convert sample result ($\bar{x}$) to a t test statistic



Consider the test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

(Assume the population is normal)

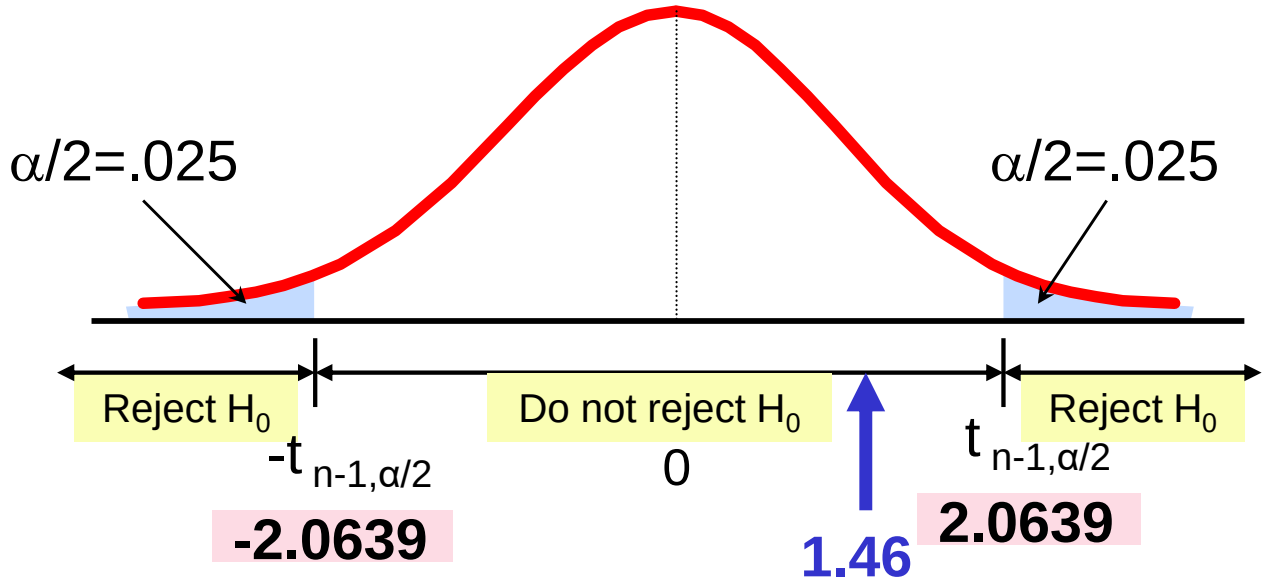
The decision rule is:

$$\text{Reject } H_0 \text{ if } t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{n-1,\alpha}$$

t-Test of Hypothesis for the Mean: Example

H_0: $\mu = 168$

H_1: $\mu \neq 168$



- $\alpha = 0.05$
- $n = 25$
- σ is unknown, so use a t statistic
- Critical Value:

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

$t_{24,.025} = \pm 2.0639$

Do not reject H_0: not sufficient evidence that true mean cost is different than \$168

t-Test for the mean in R

```
> mpg = c(11.4,13.1,14.7,14.7,15.0,15.5,15.6,15.9,16.0,16.8)
> t.test(mpg, mu = 17, alt="less")
One Sample t-test
data:  mpg
t = -4.285, df = 9, p-value = 0.001018
alternative hypothesis:  true mean is less than 17
95 percent confidence interval:
-Inf 15.78
sample estimates:
mean of x
14.87
```

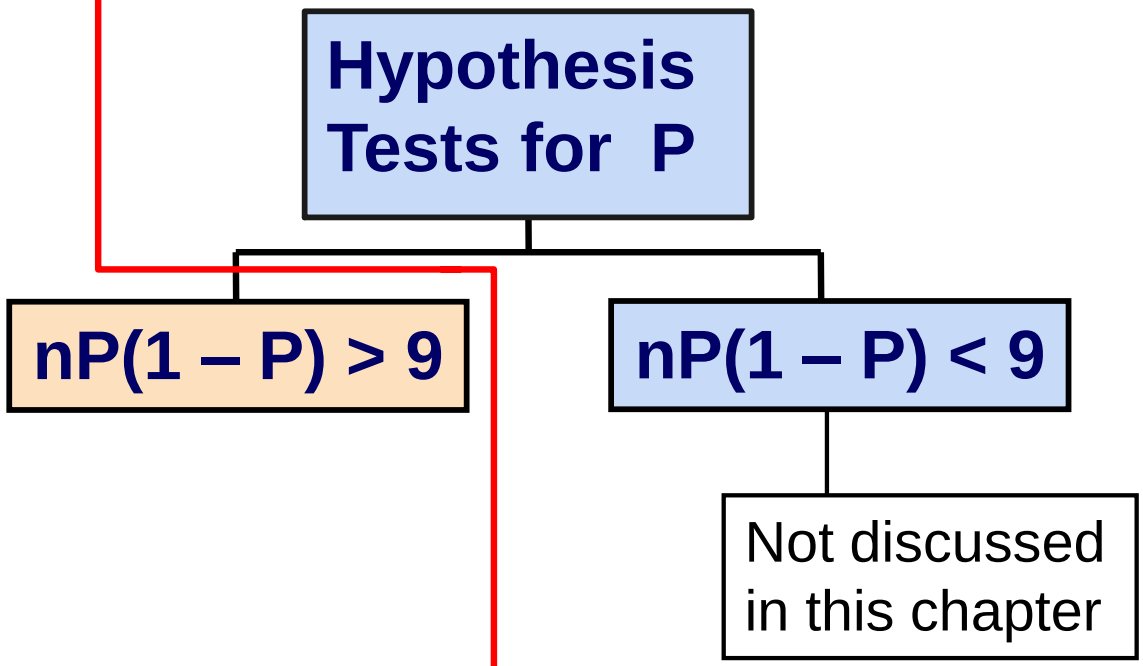
Sample Proportions

- Sample proportion in the success category is denoted by $\hat{p}$
  - $$\hat{p} = \frac{x}{n} = \frac{\text{number of successes in sample}}{\text{sample size}}$$
- When $nP(1 - P) > 9$, $\hat{p}$ can be approximated by a normal distribution with mean and standard deviation
  - $$\mu_{\hat{p}} = P$$
  - $$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$$

Hypothesis Tests for Proportions

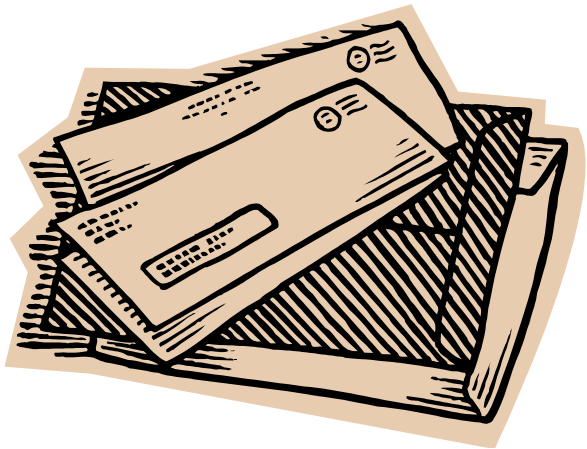
- The sampling distribution of $\hat{p}$ is approximately normal, so the test statistic is a z value:

$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$



Example: Z Test for Proportion

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the $\alpha = .05$ significance level.



Check:

Our approximation for P is

$\hat{p} = 25/500 = .05$

$nP(1 - P) = (500)(.05)(.95)$
$= 23.75 > 9$

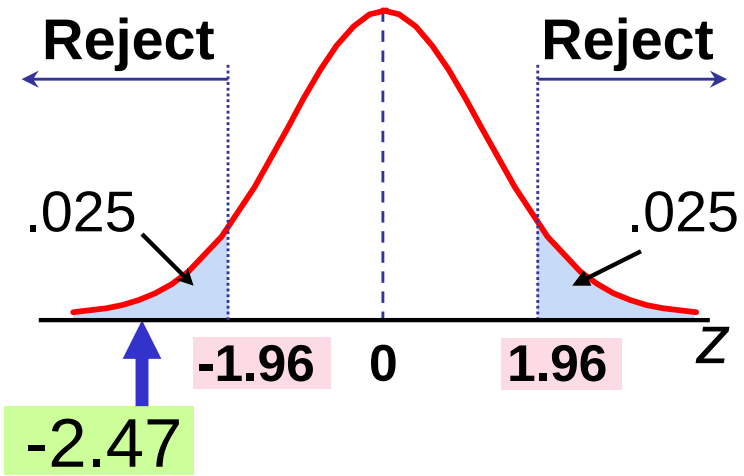
Example: Z Test for Proportion (2)

H_0: P = .08
H_1: P ≠ .08

α = .05

n = 500, $\hat{p}$ = .05

Critical Values: ± 1.96



Test Statistic:

$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{.05 - .08}{\sqrt{\frac{.08(1-.08)}{500}}} = -2.47$$

Decision:

Reject H_0 at α = .05

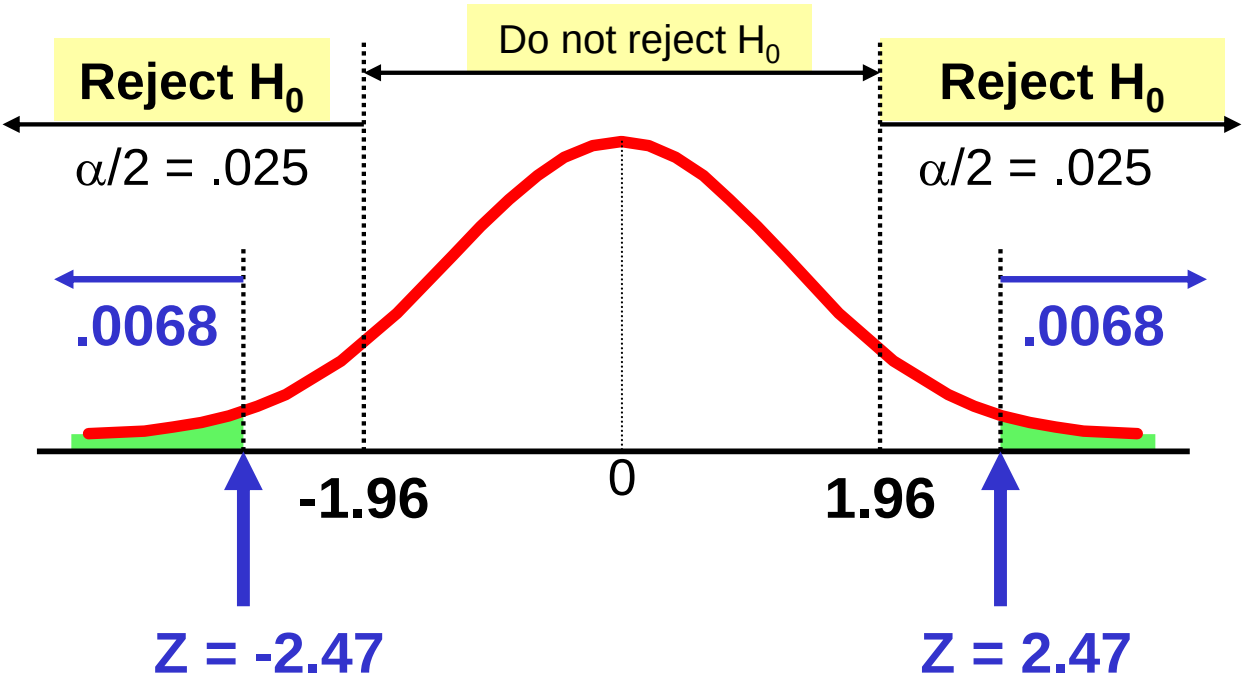
Conclusion:

There is sufficient evidence to reject the company's claim of 8% response rate.

p-Value Solution

Calculate the p-value and compare to α

(For a two sided test the p-value is always two sided)



p-value = .0136:

$$P(Z \leq -2.47) + P(Z \geq 2.47)$$

$$= 2(.0068) = 0.0136$$

Reject H_0 since p-value = .0136 < α = .05

Test for Proportion in R

- One-tail test:

```
> prop.test(x=5850, n=50000, p=.113, alt="greater")
1-sample proportions test with continuity correction
data:  5850 out of 50000, null probability 0.113
X-squared = 7.942, df = 1, p-value = 0.002415
alternative hypothesis:  true p is greater than 0.113
95 percent confidence interval:
0.1146 1.0000
sample estimates:
P
0.117
```

- Two-tail test:

```
> prop.test(x=5850, n=50000, p=.113, alt="two.sided")
...
X-squared=7.942, df=1, p-value=0.004831
```

Matched Pairs

Matched Pairs

Tests Means of 2 Related Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use difference between paired values:

$$d_i = x_i - y_i$$

- Assumptions:
  - Both Populations Are Normally Distributed

Test Statistic: Matched Pairs

Matched Pairs

The test statistic for the mean difference is a t value, with n – 1 degrees of freedom:

$$t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$$

Where

D_0 = hypothesized mean difference

s_d = sample standard dev. of differences

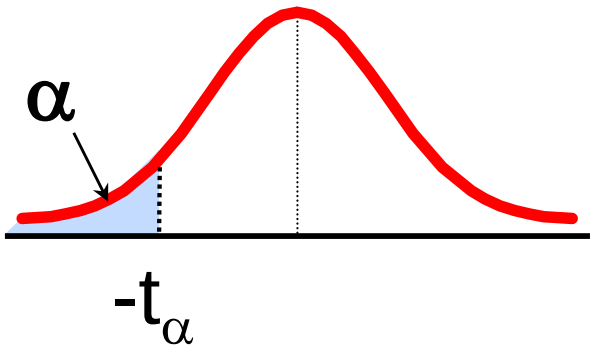
n = the sample size (number of pairs)

Decision Rules: Matched Pairs

Paired Samples

Lower-tail test:

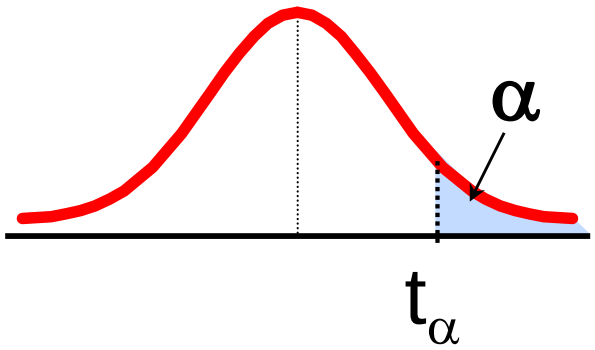
$H_0: \mu_x - \mu_y \geq 0$
$H_1: \mu_x - \mu_y < 0$



Reject H_0 if $t < -t_{n-1,\,\alpha}$

Upper-tail test:

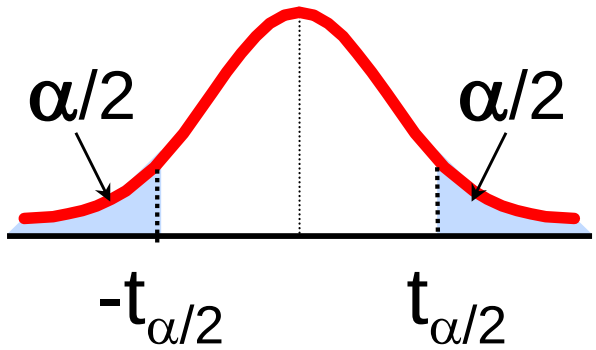
$H_0: \mu_x - \mu_y \leq 0$
$H_1: \mu_x - \mu_y > 0$



Reject H_0 if $t > t_{n-1,\,\alpha}$

Two-tail test:

$H_0: \mu_x - \mu_y = 0$
$H_1: \mu_x - \mu_y \neq 0$



Reject H_0 if $t < -t_{n-1\,,\,\alpha/2}$ or $t > t_{n-1\,,\,\alpha/2}$

Where $t = \dfrac{\bar{d} - D_0}{\dfrac{s_d}{\sqrt{n}}}$ has $n - 1$ d.f.

Matched Pairs: Example

- Assume you send your salespeople to a “customer service” training workshop. Has the training made a difference in the number of complaints? You collect the following data:

Salesperson	Number of Complaints: Before (1)	After (2)	(2) - (1) Difference, d_i
C.B.	6	4	- 2
T.F.	20	6	-14
M.H.	3	2	- 1
R.K.	0	0	0
M.O.	4	0	- 4
			-21

$$\bar{d} = \frac{\sum d_i}{n}$$

$$= -4.2$$

$$S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

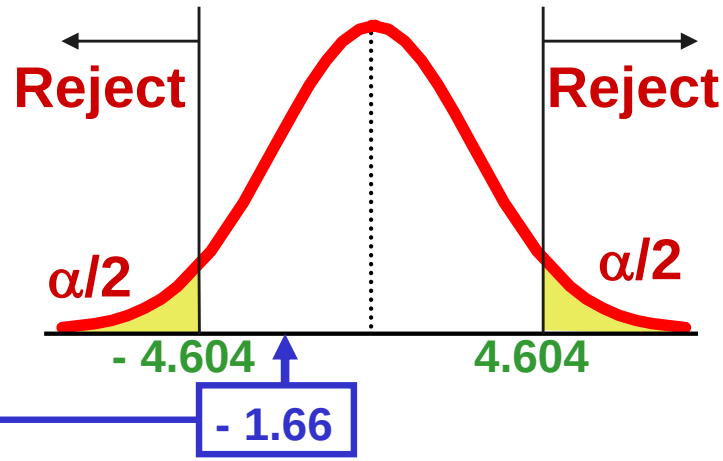
$$= 5.67$$

Matched Pairs: Solution

- Has the training made a difference in the number of complaints (at the $\alpha = 0.01$ level)?

$$H_0: \mu_x - \mu_y = 0$$
$$H_1: \mu_x - \mu_y \neq 0$$

$\alpha = .01$ $\quad \bar{d} = -4.2$



Critical Value = ± 4.604

d.f. = n - 1 = 4

Test Statistic:

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{5}} = -1.66$$

Decision: Do not reject H_0

(t stat is not in the reject region)

Conclusion: There is not a significant change in the number of complaints.

Matched Pairs Test in R

```
> Finasteride = c(5,3,5,6,4,4,7,4,3)
> placebo = c(2,3,2,4,2,2,3,4,2)
> t.test(Finasteride, placebo, paired=TRUE, alt="two.sided")
Paired t-test
data:  Finasteride and placebo
t=4.154, df=8, p-value=0.003192
alternative hypothesis:  true difference in means is not
equal to a
95 percent confidence interval:
0.8403 2.9375
sample estimates:
mean of the differences
1.889
```

- One-tail test:

```
> pre = c(77, 56, 64, 60, 57, 53, 72, 62, 65, 66)
> post = c(88, 74, 83, 68, 58, 50, 67, 64, 74, 60)
> t.test(pre, post,var.equal=TRUE, alt="less")
...
t = -1.248, df = 18, p-value = 0.1139
```

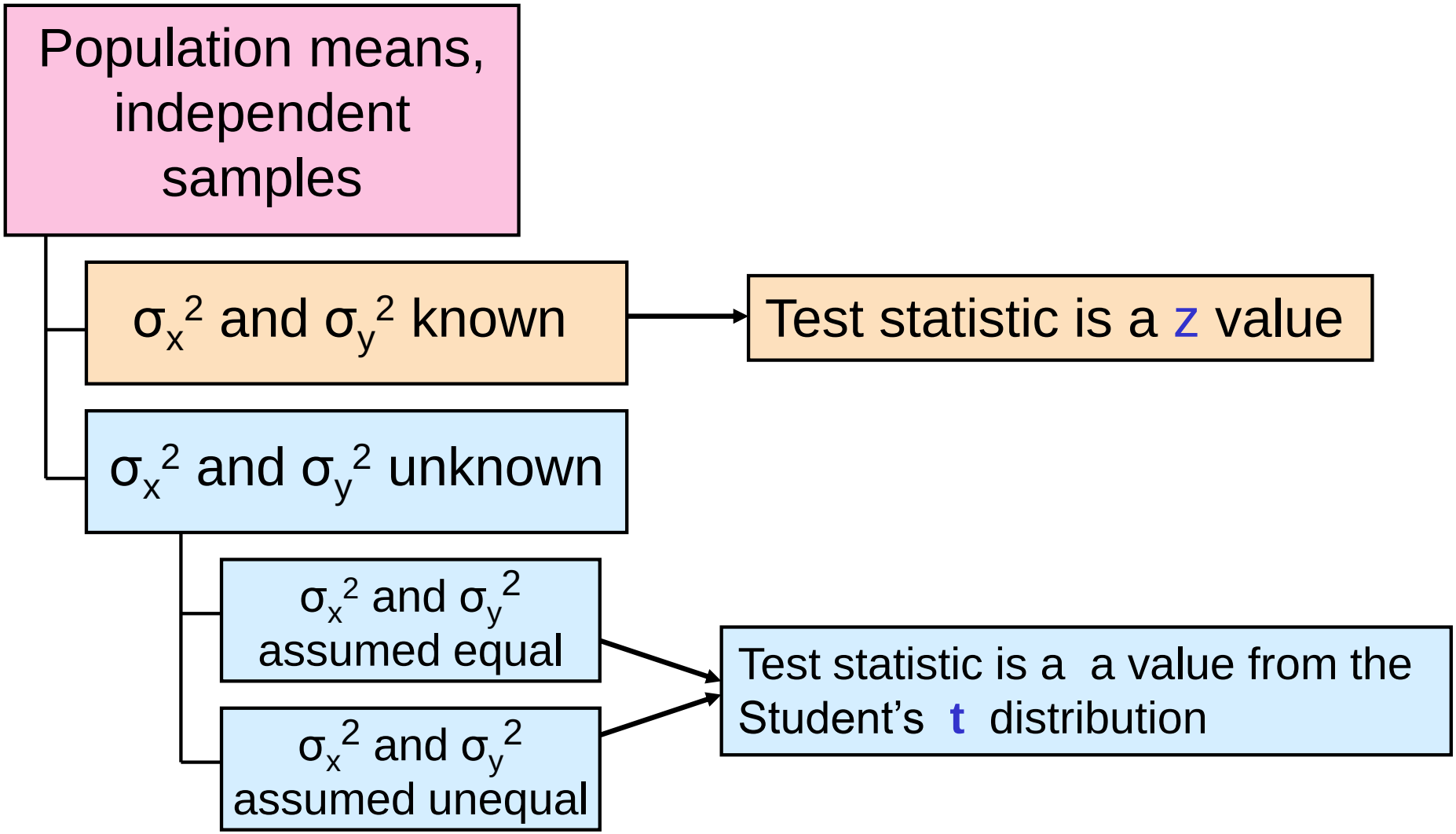
Difference Between Two Means

Population means, independent samples

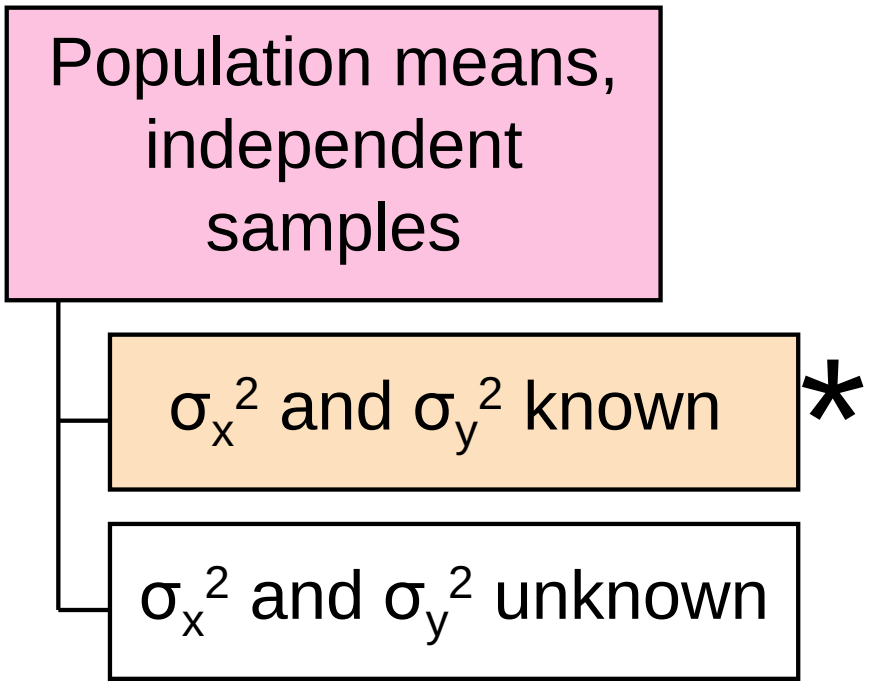
Goal: Form a confidence interval for the difference between two population means, $\mu_x - \mu_y$

- Different data sources
  - Unrelated
  - Independent
    - Sample selected from one population has no effect on the sample selected from the other population

Difference Between Two Means (2)



σ_x^2 and σ_y^2 known



Assumptions:

- Samples are randomly and independently drawn
- both population distributions are normal
- Population variances are known

σ_x^2 and σ_y^2 known (2)

- Population means, independent samples
  - σ_x^2 and σ_y^2 known *
  - σ_x^2 and σ_y^2 unknown

When σ_x^2 and σ_y^2 are known and both populations are normal, the variance of $\overline{X} - \overline{Y}$ is

$$\sigma^2_{\overline{X}-\overline{Y}} = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

…and the random variable

$$Z = \frac{(\overline{x}-\overline{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_x^2}{n_X}+\frac{\sigma_y^2}{n_Y}}}$$

has a standard normal distribution

Test statistic: σ_x^2 and σ_y^2 known

- Population means, independent samples
  - σ_x^2 and σ_y^2 known *
  - σ_x^2 and σ_y^2 unknown

The test statistic for $\mu_x - \mu_y$ is:

$$z = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Hypothesis Tests for Two Population Means

Two Population Means, Independent Samples

Lower-tail test:

$H_0: \mu_x \geq \mu_y$
$H_1: \mu_x < \mu_y$

i.e.,

$H_0: \mu_x - \mu_y \geq 0$
$H_1: \mu_x - \mu_y < 0$

Upper-tail test:

$H_0: \mu_x \leq \mu_y$
$H_1: \mu_x > \mu_y$

i.e.,

$H_0: \mu_x - \mu_y \leq 0$
$H_1: \mu_x - \mu_y > 0$

Two-tail test:

$H_0: \mu_x = \mu_y$
$H_1: \mu_x \neq \mu_y$

i.e.,

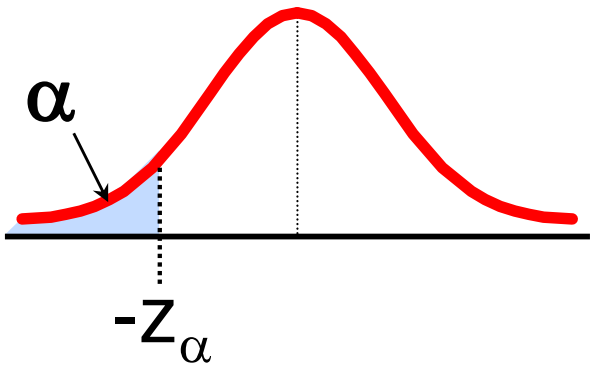
$H_0: \mu_x - \mu_y = 0$
$H_1: \mu_x - \mu_y \neq 0$

Decision Rules

Two Population Means, Independent Samples, Variances Known

Lower-tail test:

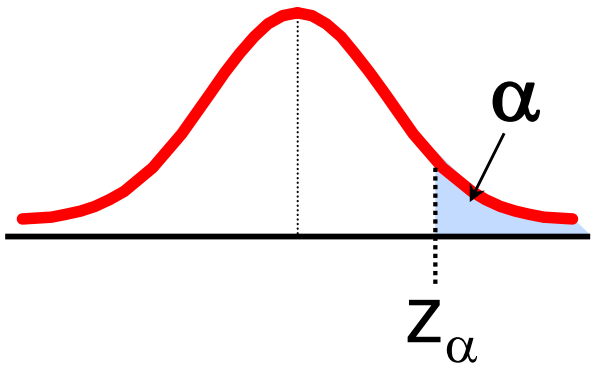
$H_0: \mu_x - \mu_y \geq 0$
$H_1: \mu_x - \mu_y < 0$



Reject H_0 if $z < -z_\alpha$

Upper-tail test:

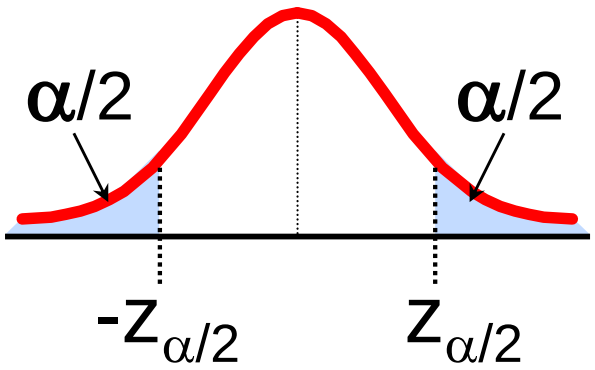
$H_0: \mu_x - \mu_y \leq 0$
$H_1: \mu_x - \mu_y > 0$



Reject H_0 if $z > z_\alpha$

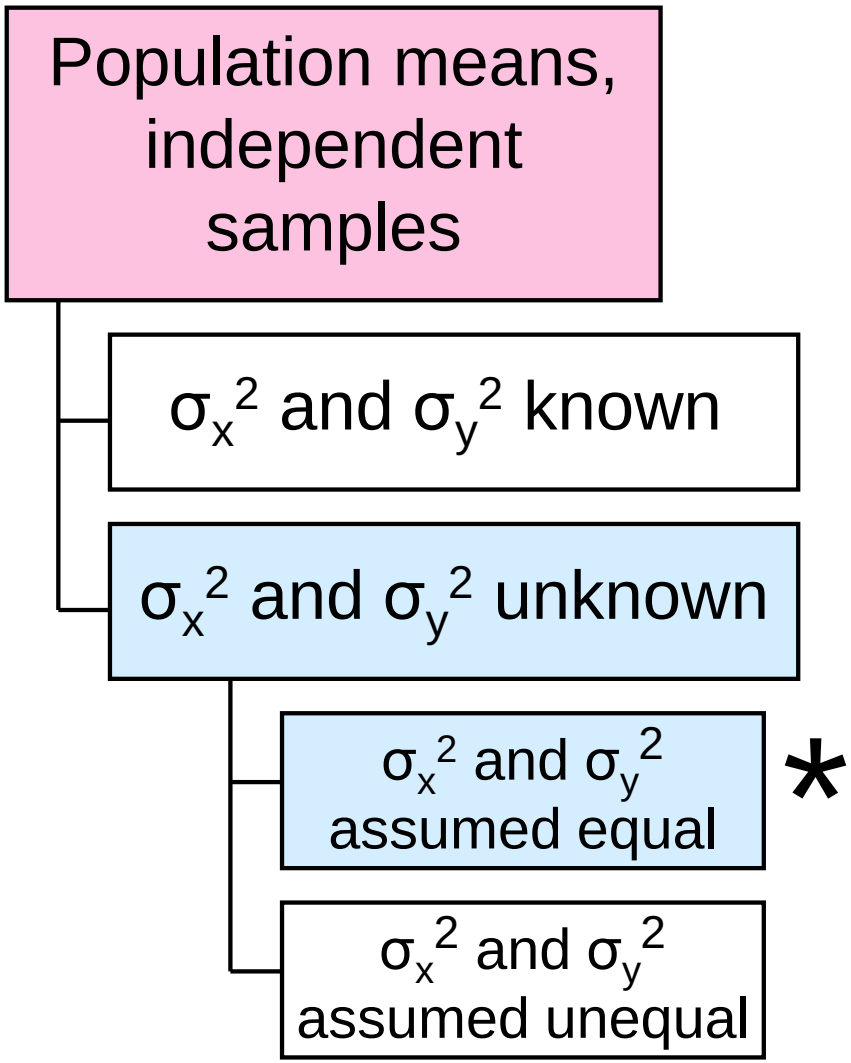
Two-tail test:

$H_0: \mu_x - \mu_y = 0$
$H_1: \mu_x - \mu_y \neq 0$



Reject H_0 if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

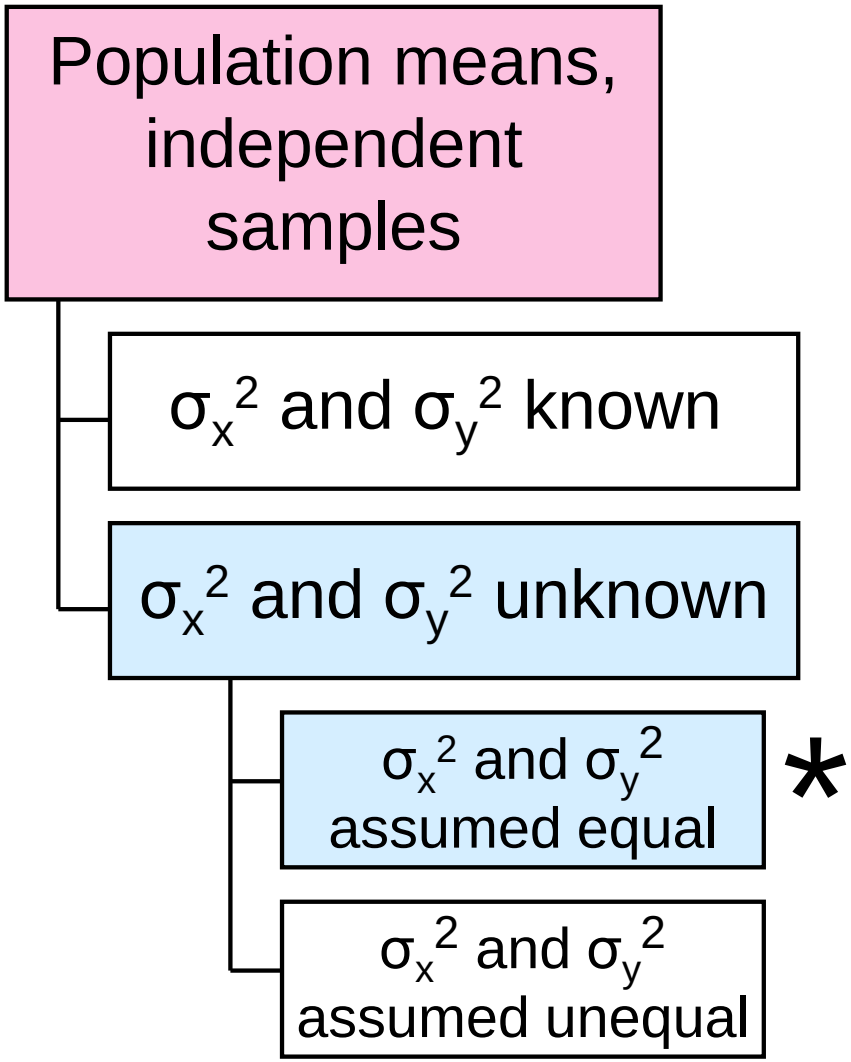
σ_x^2 and σ_y^2 unknown, assumed equal



Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown but assumed equal

σ_x^2 and σ_y^2 unknown, assumed equal (2)



Forming interval estimates:

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ
- use a t value with $(n_x + n_y - 2)$ degrees of freedom

σ_x^2 and σ_y^2 unknown, assumed equal (3)

- σ_x^2 and σ_y^2 unknown
  - σ_x^2 and σ_y^2 assumed equal *
  - σ_x^2 and σ_y^2 assumed unequal

The test statistic for $\mu_x - \mu_y$ is:

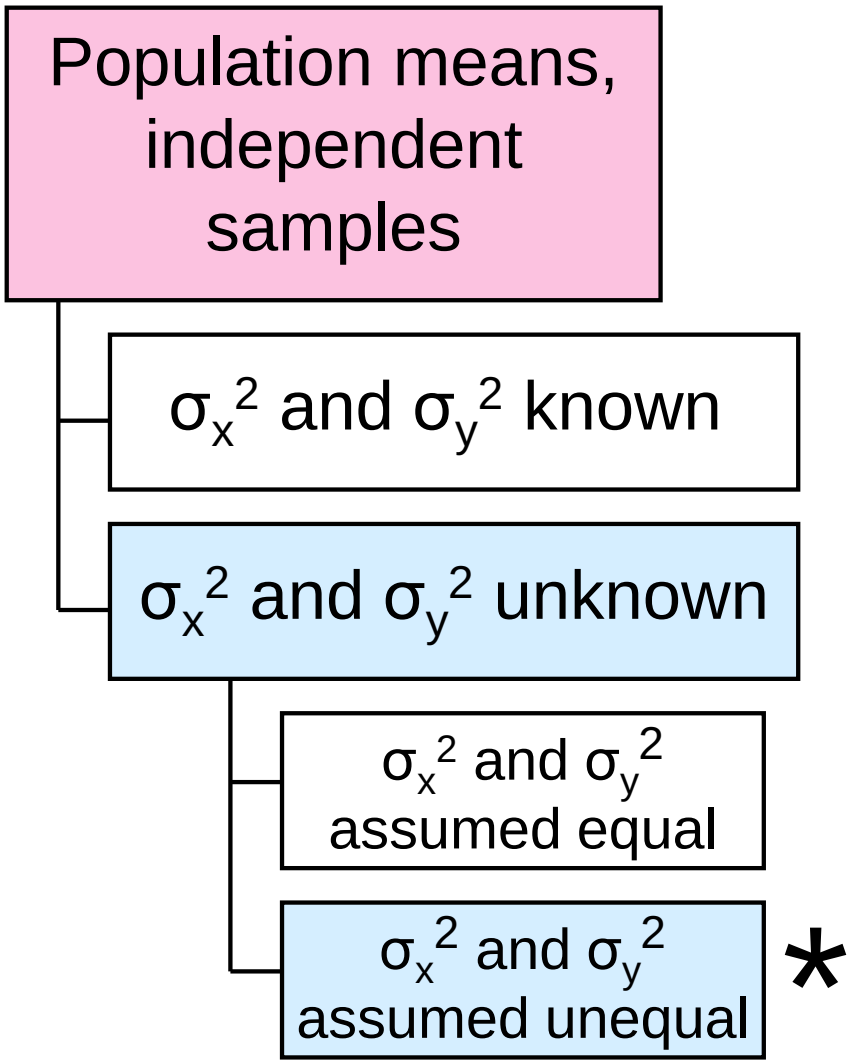
$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

Where t has $(n_1 + n_2 - 2)$ d.f.,

and

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

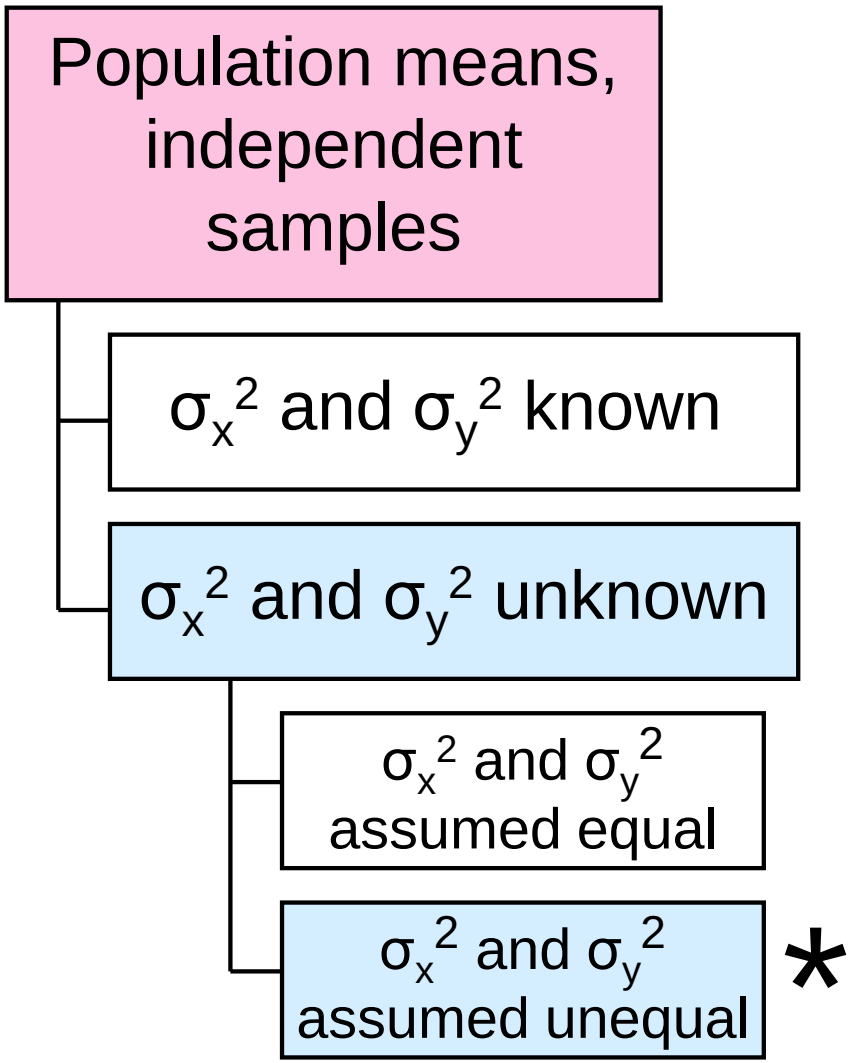
Test statistic: σ_x^2 and σ_y^2 unknown, assumed equal



Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown and assumed unequal

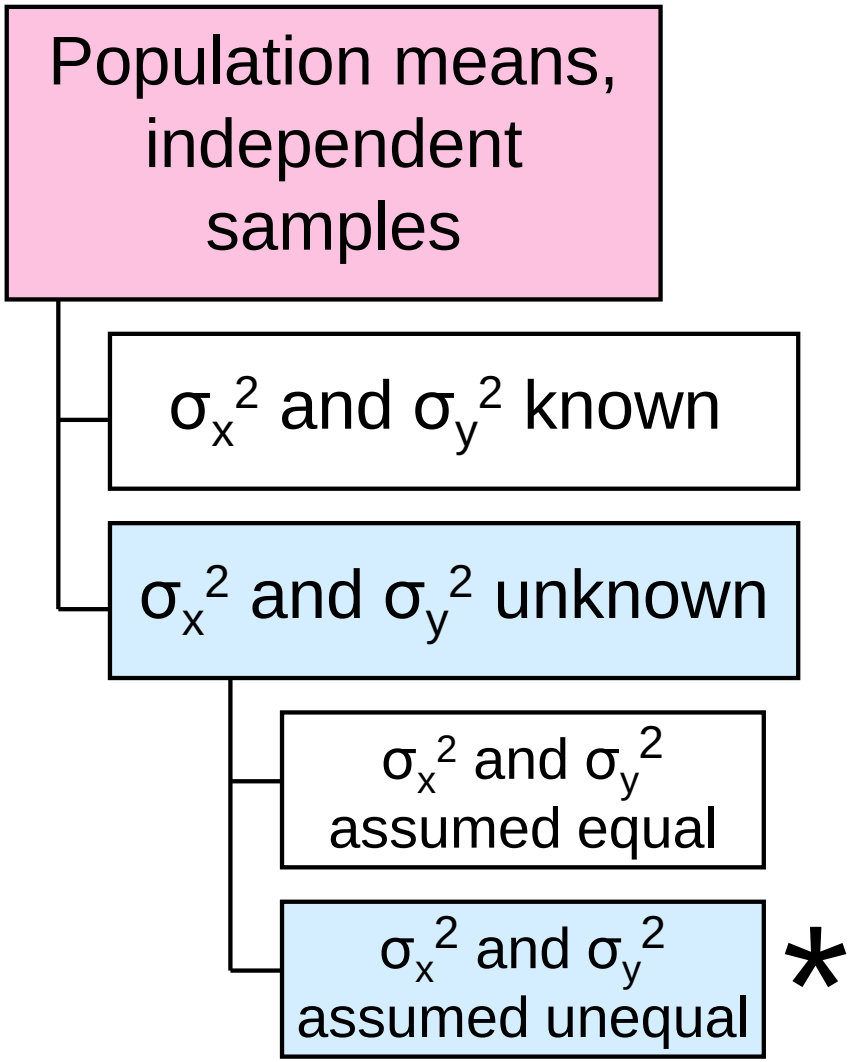
σ_x^2 and σ_y^2 unknown, assumed unequal



Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown and assumed unequal

σ_x^2 and σ_y^2 unknown, assumed unequal (2)



Forming interval estimates:

- The population variances are assumed unequal, so a pooled variance is not appropriate
- use a t value with ν degrees of freedom, where

$$\nu = \frac{\left[\left(\frac{s_x^2}{n_x}\right)+\left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$$

Test statistic: σ_x^2 and σ_y^2 unknown, assumed unequal

- σ_x^2 and σ_y^2 unknown
  - σ_x^2 and σ_y^2 assumed equal
  - σ_x^2 and σ_y^2 assumed unequal *

The test statistic for $\mu_x - \mu_y$ is:

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{\sigma_x^2}{n_X} + \frac{\sigma_y^2}{n_Y}}}$$

Where t has ν degrees of freedom:

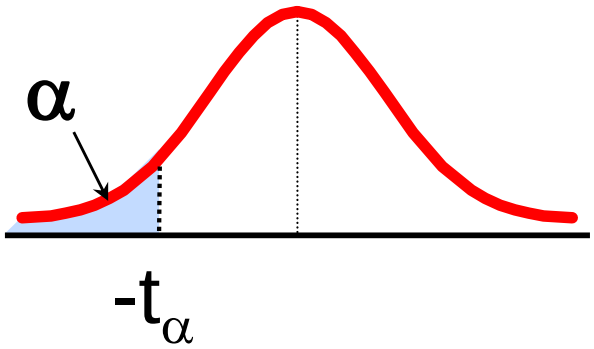
$$\nu = \frac{\left[\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2 / (n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2 / (n_y - 1)}$$

Decision Rules

Two Population Means, Independent Samples, Variances Unknown

Lower-tail test:

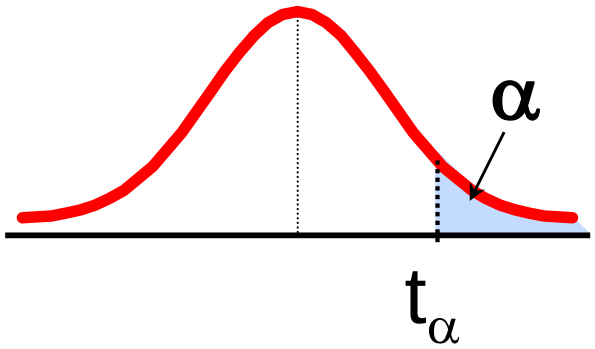
$H_0: \mu_x - \mu_y \geq 0$
$H_1: \mu_x - \mu_y < 0$



Reject H_0 if $t < -t_{n-1,\,\alpha}$

Upper-tail test:

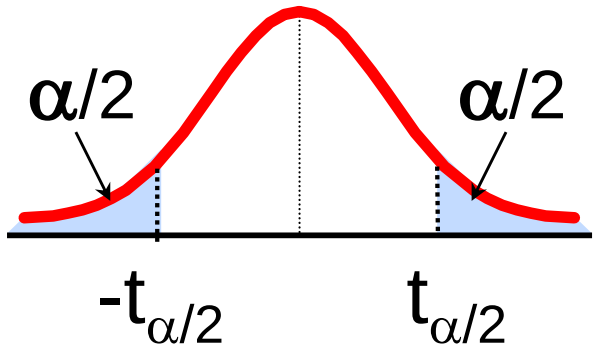
$H_0: \mu_x - \mu_y \leq 0$
$H_1: \mu_x - \mu_y > 0$



Reject H_0 if $t > t_{n-1,\,\alpha}$

Two-tail test:

$H_0: \mu_x - \mu_y = 0$
$H_1: \mu_x - \mu_y \neq 0$



Reject H_0 if $t < -t_{n-1\,,\,\alpha/2}$ or $t > t_{n-1\,,\,\alpha/2}$

Where t has n - 1 d.f.

Pooled Variance t-Test: Example

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number	**21**	**25**
Sample mean	**3.27**	**2.53**
Sample std dev	**1.30**	**1.16**

Assuming both populations are approximately normal with equal variances, is there a difference in average yield ($\alpha = 0.05$)?

Calculating the Test Statistic

The test statistic is:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\left(3.27 - 2.53\right) - 0}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \boxed{2.040}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{(n_1 - 1) + (n_2 - 1)} = \frac{(21-1)1.30^2 + (25-1)1.16^2}{(21\text{-}1) + (25 - 1)} = 1.5021$$

Solution

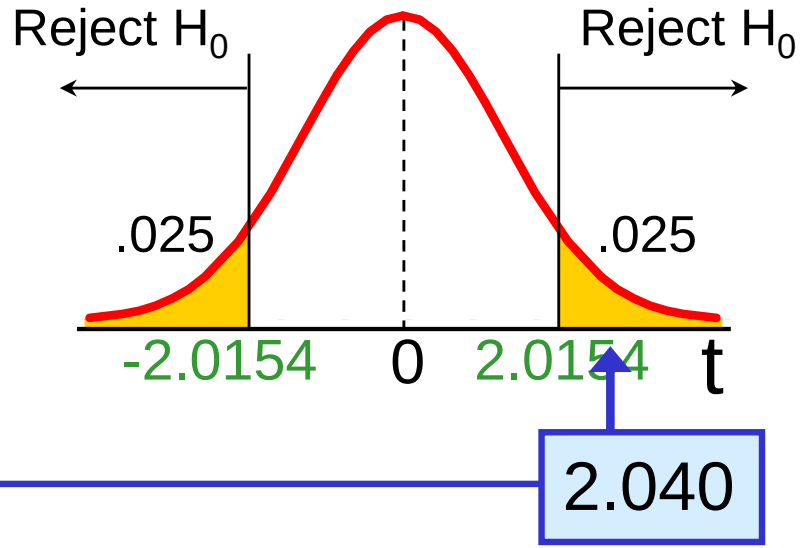
$H_0: \mu_1 - \mu_2 = 0$ i.e. $(\mu_1 = \mu_2)$

$H_1: \mu_1 - \mu_2 \neq 0$ i.e. $(\mu_1 \neq \mu_2)$

$\alpha = 0.05$

$df = 21 + 25 - 2 = 44$

Critical Values: t = ± 2.0154



Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = 2.040$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.

Two Population Proportions

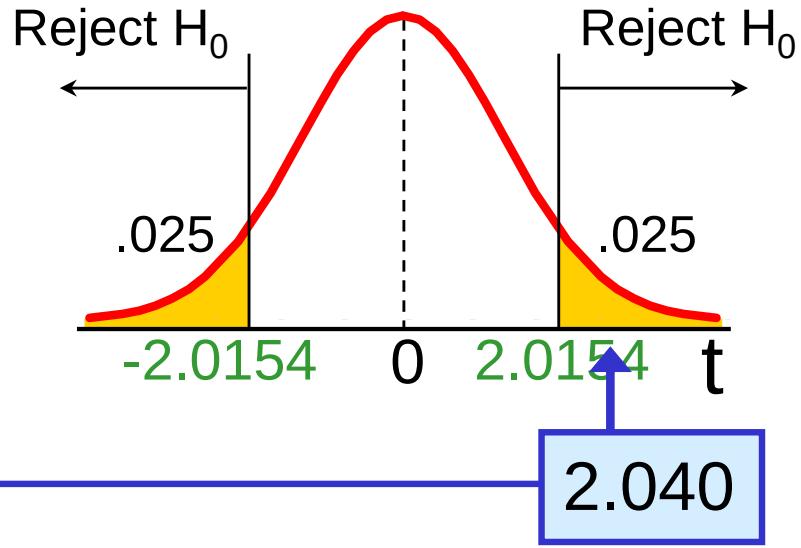
$H_0: \mu_1 - \mu_2 = 0$ i.e. $(\mu_1 = \mu_2)$

$H_1: \mu_1 - \mu_2 \neq 0$ i.e. $(\mu_1 \neq \mu_2)$

$\alpha = 0.05$

$df = 21 + 25 - 2 = 44$

Critical Values: t = ± 2.0154



Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = 2.040$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.

Two Population Proportions (2)

Population proportions

Goal: Test hypotheses for the difference between two population proportions, $P_x - P_y$

Assumptions:

Both sample sizes are large,

$$nP(1 - P) > 9$$

t-Test for the difference between two means (assumed equal) in R

```
> x = c(284, 279, 289, 292, 287, 295, 285, 279, 306, 298)
> y = c(298, 307, 297, 279, 291, 335, 299, 300, 306, 291)
> t.test(x,y,var.equal=TRUE)
Two Sample t-test
data:  x and y
t = -2.034, df = 18, p-value = 0.05696
alternative hypothesis:  true difference in means is not equal
to 0
```

t-Test for the difference between two means (assumed unequal) in R

```
> t.test(x,y)
Welch Two Sample t-test
data:  x and y
t = -2.034, df = 14.51, p-value = 0.06065
alternative hypothesis:  true difference in means is not
equal to 0
95 percent confidence interval:
-22.3557 0.5557
sample estimates:
mean of x mean of y
289.4 300.3
```

Test Statistic for Two Population Proportions

Population proportions

- The random variable

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}$$

is approximately normally distributed

Decision Rules: Proportions

Population proportions

The test statistic for

$H_0: P_x - P_y = 0$

is a z value:

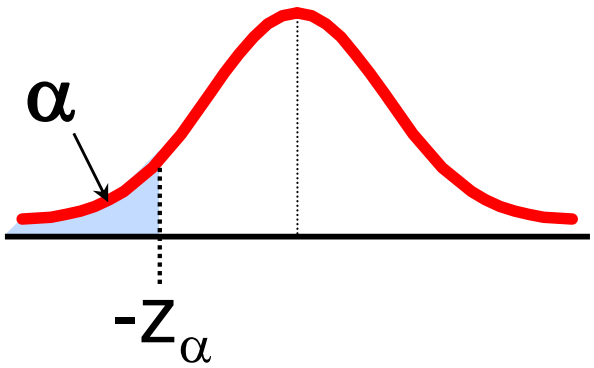
$$z = \frac{\left(\hat{p}_x - \hat{p}_y\right)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}$$

Where $\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$

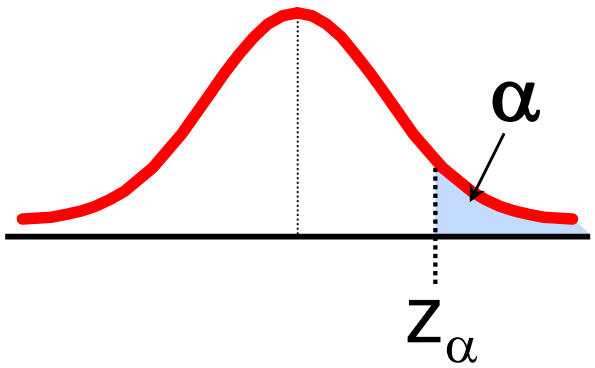
Example: Two Population Proportions

Population proportions

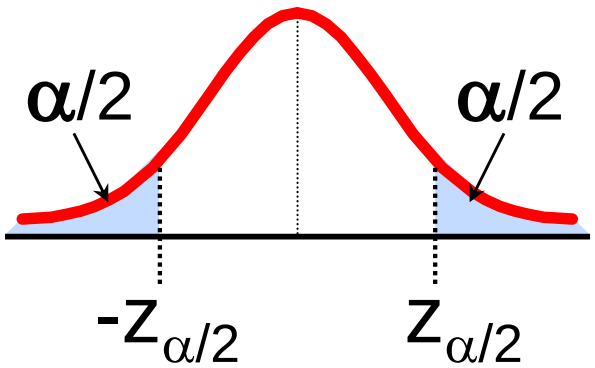
Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: p_x - p_y \geq 0$ $H_1: p_x - p_y < 0$	$H_0: p_x - p_y \leq 0$ $H_1: p_x - p_y > 0$	$H_0: p_x - p_y = 0$ $H_1: p_x - p_y \neq 0$



Reject H_0 if $z < -z_\alpha$



Reject H_0 if $z > z_\alpha$



Reject H_0 if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Example: Two Population Proportions (2)

Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A?

- In a random sample, 36 of 72 men and 31 of 50 women indicated they would vote Yes
- Test at the .05 level of significance

Example: Two Population Proportions (3)

- The hypothesis test is:

H_0: $P_M - P_W = 0$ (the two proportions are equal)

H_1: $P_M - P_W \neq 0$ (there is a significant difference between proportions)

- The sample proportions are:
  - Men: $\hat{p}_M = 36/72 = .50$
  - Women: $\hat{p}_W = 31/50 = .62$
- The estimate for the common overall proportion is:

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y} = \frac{72(36/72) + 50(31/50)}{72 + 50} = \frac{67}{122} = .549$$

Test for Two Population Proportions in R

```
> phat = c(.121, .117) # the sample proportions
> n = c(50000, 60000) # the sample sizes
> n*phat # the counts
[1] 5850 7260
> prop.test(n*phat,n,alt="less")
2-sample test for equality of proportions with
continuity correction
data:  n * phat out of n
X-squared = 4.119, df = 1, p-value = 0.02121
alternative hypothesis:  less
95 percent confidence interval:
-1.0000000 -0.0007589
sample estimates:
prop 1 prop 2
0.117 0.121
```

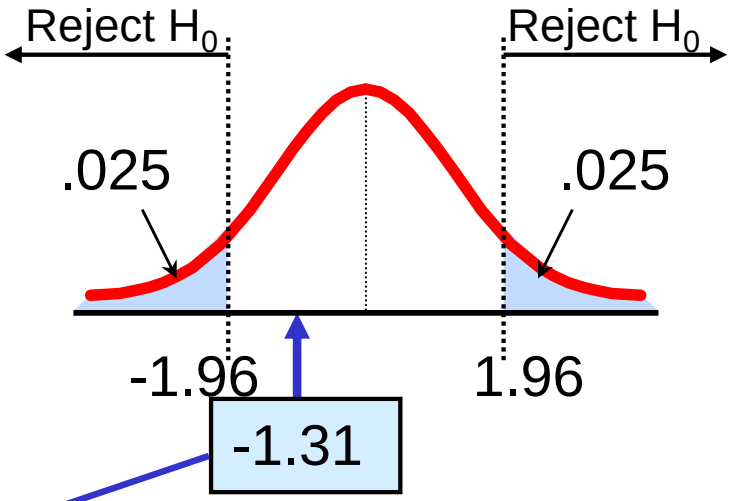
Hypothesis Tests for Population Variance

The test statistic for $P_M - P_W = 0$ is:

$$z = \frac{\left(\hat{p}_M - \hat{p}_W\right)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}}$$

$$= \frac{(.50 - .62)}{\sqrt{\left(\frac{.549(1-.549)}{72} + \frac{.549(1-.549)}{50}\right)}}$$

$$= \boxed{-1.31}$$



Critical Values = ±1.96
For α = .05

Decision: Do not reject H_0

Conclusion: There is not significant evidence of a difference between men and women in proportions who will vote yes.

Test Statistic

Population Variance

- Goal: Test hypotheses about the population variance, σ^2
- If the population is normally distributed,

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with $(n - 1)$ degrees of freedom

Decision Rules: Variance

Population Variance

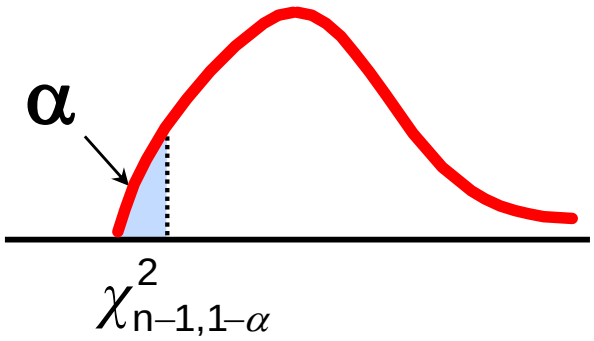
The test statistic for hypothesis tests about one population variance is

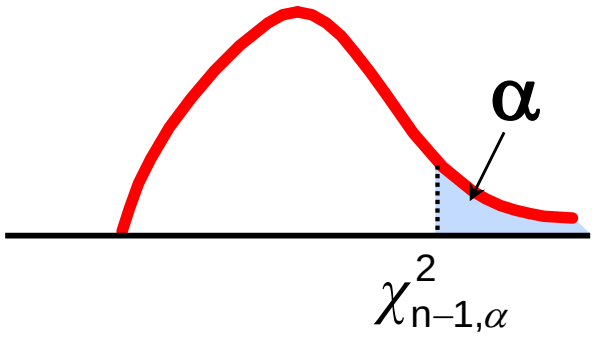
$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2}$$

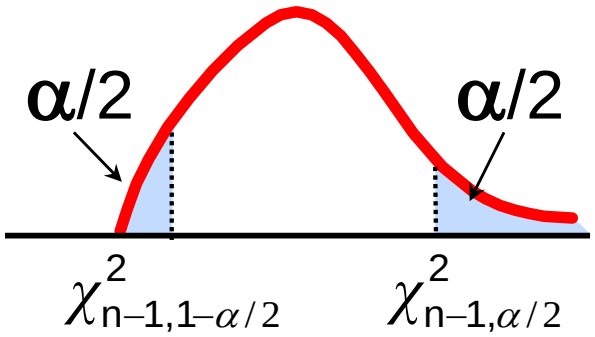
Hypothesis Tests for Two Variances

Population variance

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \sigma^2 \geq \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 < \sigma_0^2$	$H_1: \sigma^2 > \sigma_0^2$	$H_1: \sigma^2 \neq \sigma_0^2$





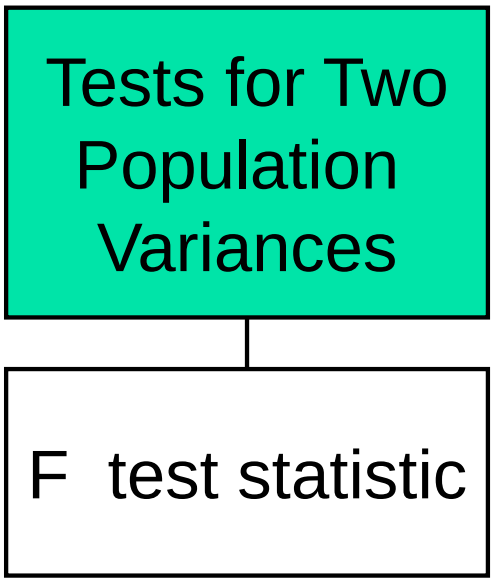


Reject H_0 if
$\chi^2_{n-1} < \chi^2_{n-1,1-\alpha}$

Reject H_0 if
$\chi^2_{n-1} > \chi^2_{n-1,\alpha}$

Reject H_0 if
$\chi^2_{n-1} > \chi^2_{n-1,\alpha/2}$
or $\chi^2_{n-1} < \chi^2_{n-1,1-\alpha/2}$

Hypothesis Tests for Two Variances (2)



- Goal: Test hypotheses about two population variances

$H_0: \sigma_x^2 \geq \sigma_y^2$
$H_1: \sigma_x^2 < \sigma_y^2$

Lower-tail test

$H_0: \sigma_x^2 \leq \sigma_y^2$
$H_1: \sigma_x^2 > \sigma_y^2$

Upper-tail test

$H_0: \sigma_x^2 = \sigma_y^2$
$H_1: \sigma_x^2 \neq \sigma_y^2$

Two-tail test

The two populations are assumed to be independent and normally distributed

Test Statistic

Tests for Two Population Variances

F test statistic

The random variable

$$F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$$

Has an F distribution with $(n_x - 1)$ numerator degrees of freedom and $(n_y - 1)$ denominator degrees of freedom

Denote an F value with ν_1 numerator and ν_2 denominator degrees of freedom by

Decision Rules: Two Variances

Tests for Two Population Variances

F test statistic

The critical value for a hypothesis test about two population variances is

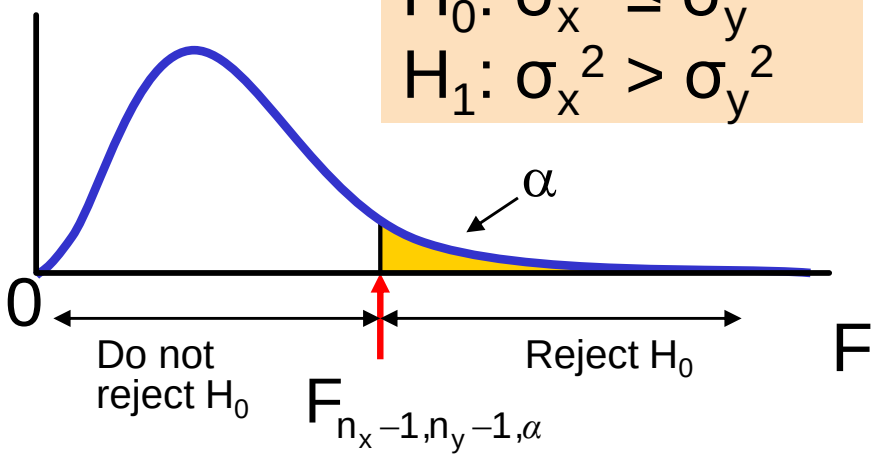
$$F = \frac{s_x^2}{s_y^2}$$

where F has $(n_x - 1)$ numerator degrees of freedom and $(n_y - 1)$ denominator degrees of freedom

Example: F Test

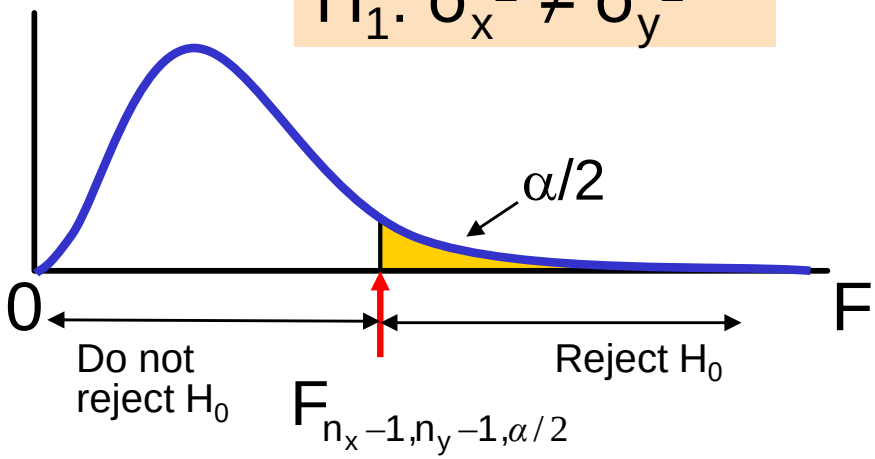
Use s_x^2 to denote the larger variance.

$H_0: \sigma_x^2 \leq \sigma_y^2$
$H_1: \sigma_x^2 > \sigma_y^2$



Reject H_0 if $F > F_{n_x-1,n_y-1,\alpha}$

$H_0: \sigma_x^2 = \sigma_y^2$
$H_1: \sigma_x^2 \neq \sigma_y^2$



- rejection region for a two-tail test is:

Reject H_0 if $F > F_{n_x-1,n_y-1,\alpha/2}$

where s_x^2 is the larger of the two sample variances

F-Test: Example

You are a financial analyst for a brokerage firm. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

	NYSE	NASDAQ
Number	**21**	**25**
Mean	**3.27**	**2.53**
Std dev	**1.30**	**1.16**

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.10$ level?

F-Test: Example (2)

- Form the hypothesis test:

$H_0: \sigma_x^2 = \sigma_y^2$ (there is no difference between variances)

$H_1: \sigma_x^2 \neq \sigma_y^2$ (there is a difference between variances)

- Find the F critical values for $\alpha = .10/2$:

Degrees of Freedom:

- Numerator

  (NYSE has the larger standard deviation):
  - $n_x - 1 = 21 - 1 = 20$ d.f.
- Denominator:
  - $n_y - 1 = 25 - 1 = 24$ d.f.

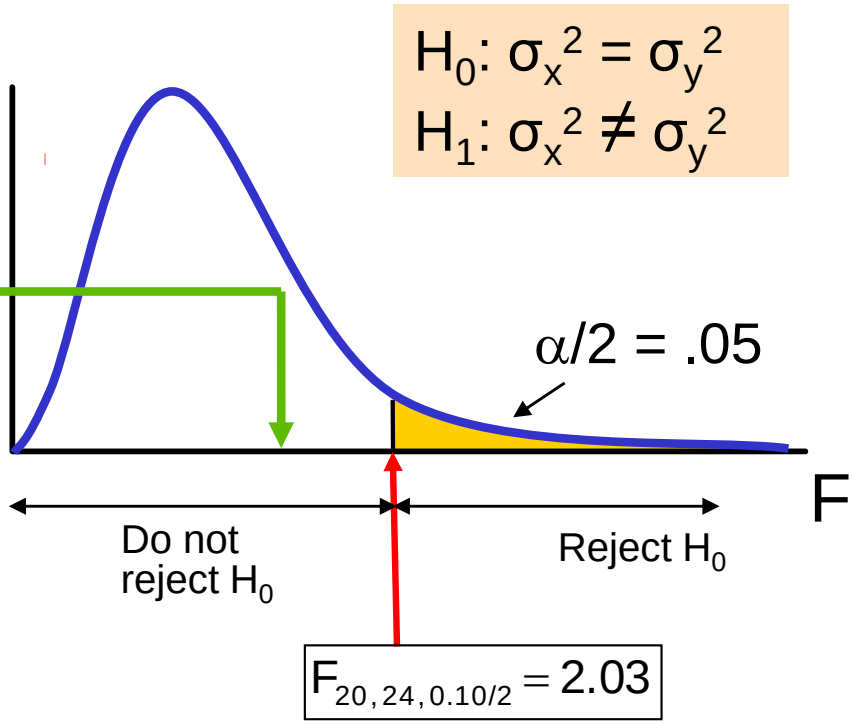
$$F_{n_x - 1, n_y - 1, \alpha/2}$$

$$= F_{20, 24, 0.10/2} = 2.03$$

F-Test: Example (3)

- The test statistic is:

$$F = \frac{s_x^2}{s_y^2} = \frac{1.30^2}{1.16^2} = 1.256$$



- F = 1.256 is not in the rejection region, so we do not reject H_0

- Conclusion: There is not sufficient evidence of a difference in variances at $\alpha = .10$

Tests for variance in R

Function `var.test()`

Chi-Square Goodness-of-Fit Test

- Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution?)
  - Sample data for 10 days per day of week:

	Sum of calls for this day:
Monday	290
Tuesday	250
Wednesday	238
Thursday	257
Friday	265
Saturday	230
Sunday	192
	$\Sigma = 1722$

Logic Behind Goodness-of-Fit Test

- If calls **are** uniformly distributed, the 1722 calls would be expected to be equally divided across the 7 days:

$$\frac{1722}{7} = 246 \quad \text{expected calls per day if uniform}$$

- Chi-Square Goodness-of-Fit Test: test to see if the sample results are consistent with the expected results

Observed vs. Expected Frequencies

	Observed O_i	Expected E_i
Monday	290	246
Tuesday	250	246
Wednesday	238	246
Thursday	257	246
Friday	265	246
Saturday	230	246
Sunday	192	246
TOTAL	1722	1722

Chi-Square Test Statistic

H_0: The distribution of calls is uniform over days of the week

H_1: The distribution of calls is not uniform

- The test statistic is

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \quad (\text{where d.f.} = K - 1)$$

where:

K = number of categories

O_i = observed frequency for category i

E_i = expected frequency for category i

The Rejection Region

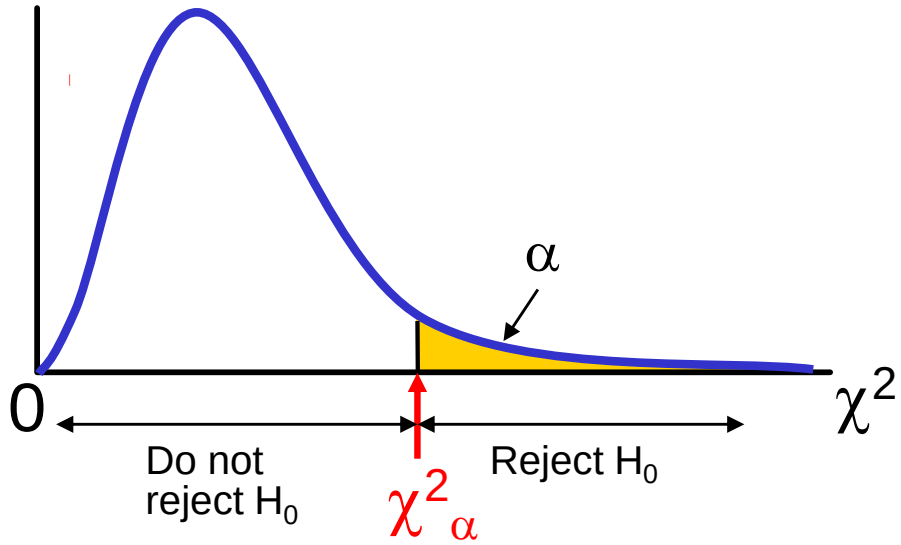
H_0: The distribution of calls is uniform over days of the week

H_1: The distribution of calls is not uniform

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

- Reject H_0 if $\chi^2 > \chi^2_\alpha$

(with k – 1 degrees of freedom)



Chi-Square Test Statistic

H_0: The distribution of calls is uniform over days of the week

H_1: The distribution of calls is not uniform

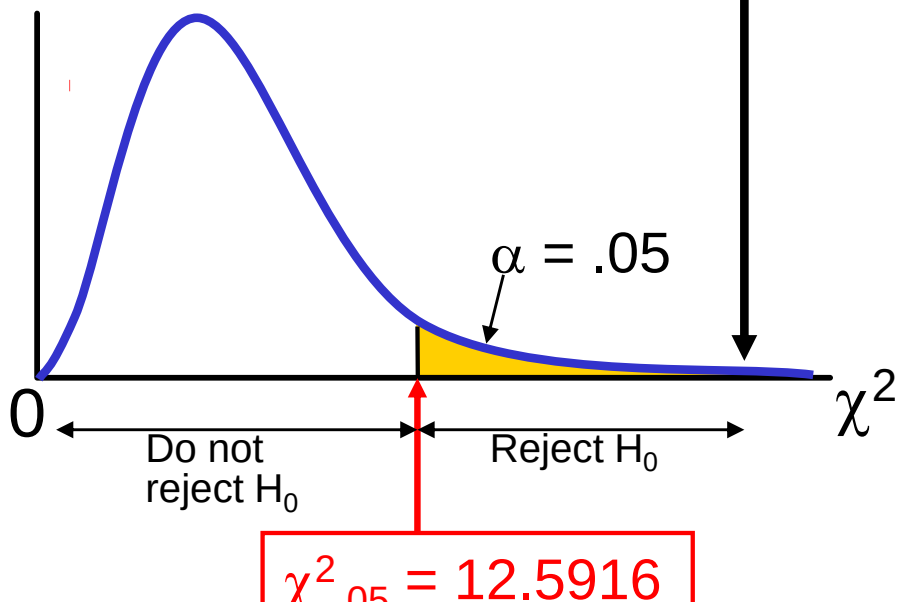
$$\chi^2 = \frac{(290-246)^2}{246} + \frac{(250-246)^2}{246} + \ldots + \frac{(192-246)^2}{246} = 23.05$$

k – 1 = 6 (7 days of the week) so use 6 degrees of freedom:

$\chi^2_{.05} = 12.5916$

Conclusion:

$\chi^2 = 23.05 > \chi^2_{\alpha} = 12.5916$ so **reject H_0** and conclude that the distribution is not uniform



Goodness-Of-Fit Test in R

```
> y = c(35,40,25)
> chisq.test(y, p=rep(1/length(da),length(da)))
Chi-squared test for given probabilities
data:  y
X-squared = 1.548, df = 2, p-value = 0.4613
```

Goodness-of-Fit Tests, Population Parameters Unknown

Idea:

- Test whether data follow a specified distribution (such as binomial, Poisson, or normal) . . .
- . . . without assuming the parameters of the distribution are known
- Use sample data to estimate the unknown population parameters

Goodness-of-Fit Tests, Population Parameters Unknown (2)

- Suppose that a null hypothesis specifies category probabilities that depend on the estimation (from the data) of m unknown population parameters
- The appropriate goodness-of-fit test is the same as in the previously section . . .

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

- . . . except that the number of degrees of freedom for the chi-square random variable is

$$\text{Degrees of Freedom} = (K - m - 1)$$

- Where K is the number of categories

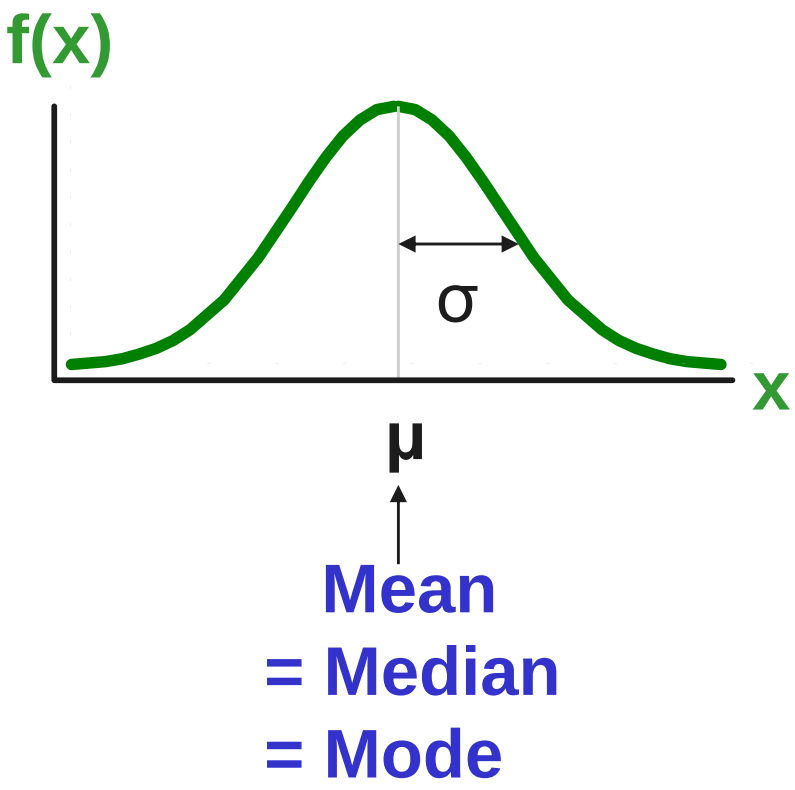
The Normal Distribution

- Bell Shaped
- Symmetrical
- Mean, Median and Mode are Equal

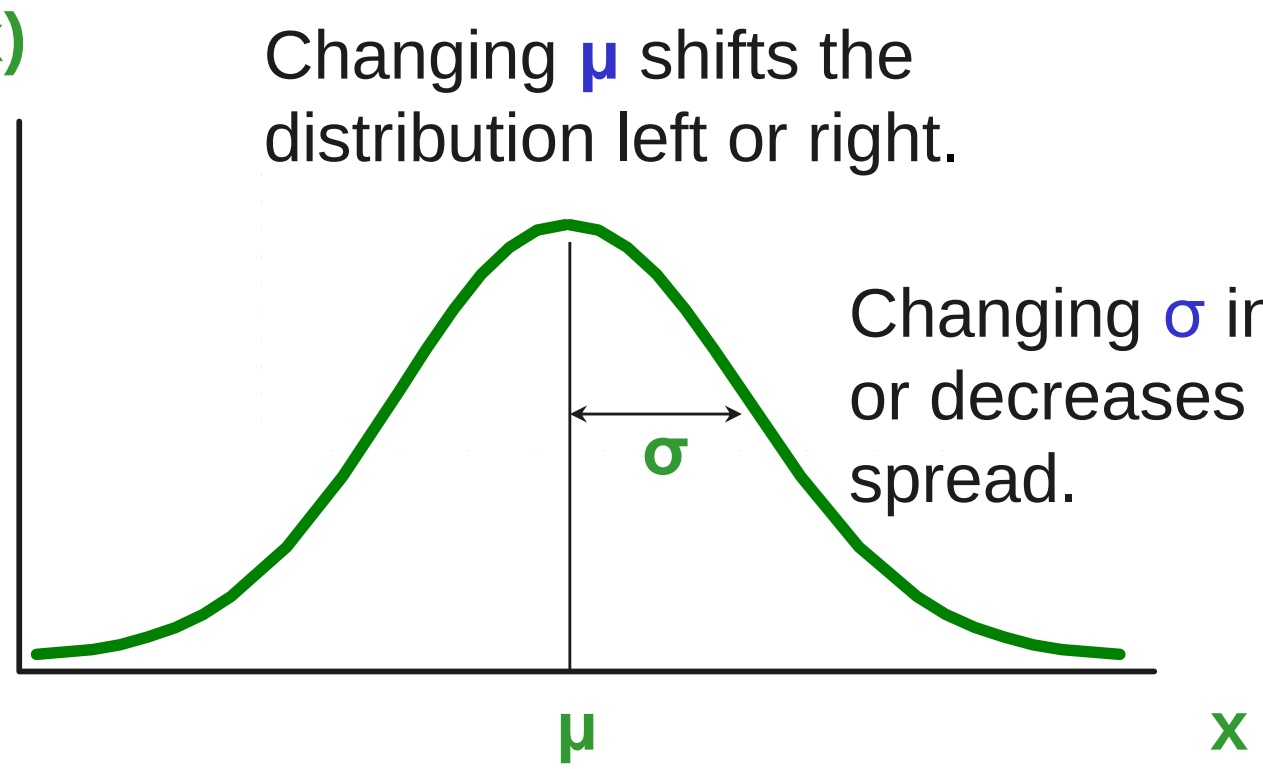
Location is determined by the mean, μ

Spread is determined by the standard deviation, σ

The random variable has an infinite theoretical range:
$+\infty$ **to** $-\infty$



The Normal Distribution Shape



Given the mean μ and variance σ we define the normal distribution using the notation

$$X \sim N(\mu, \sigma^2)$$

The Normal Probability Density Function

- The formula for the normal probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

Where e = the mathematical constant approximated by 2.71828

π = the mathematical constant approximated by 3.14159

μ = the population mean

σ = the population standard deviation

x = any value of the continuous variable, $-\infty < x < \infty$

Test of Normality

- The assumption that data follow a normal distribution is common in statistics
- Normality was assessed in prior chapters
  - Normal probability plots (Chapter 6)
  - Normal quintile plots (Chapter 8)
- Here, a chi-square test is developed

Test of Normality (2)

- Two population parameters can be estimated using sample data:

$$\text{Skewness} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^3}{ns^3}$$

$$\text{Kurtosis} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^4}{ns^4}$$

- For a normal distribution,

Skewness = 0

Kurtosis = 3

Bowman-Shelton Test for Normality

- Consider the null hypothesis that the population distribution is normal
- The Bowman-Shelton Test for Normality is based on the closeness the sample skewness to 0 and the sample kurtosis to 3
- The test statistic is

$$B = n\left[\frac{(\text{Skewness})^2}{6} + \frac{(\text{Kurtosis} - 3)^2}{24}\right]$$

- as the number of sample observations becomes very large, this statistic has a chi-square distribution with 2 degrees of freedom
- The null hypothesis is rejected for large values of the test statistic

Bowman-Shelton Test for Normality (2)

- The chi-square approximation is close only for very large sample sizes
- If the sample size is not very large, the Bowman-Shelton test statistic is compared to significance points from text Table 16.7

Sample size N	10% point	5% point	Sample size N	10% point	5% point
20	2.13	3.26	200	3.48	4.43
30	2.49	3.71	250	3.54	4.61
40	2.70	3.99	300	3.68	4.60
50	2.90	4.26	400	3.76	4.74
75	3.09	4.27	500	3.91	4.82
100	3.14	4.29	800	4.32	5.46
125	3.31	4.34	∞	4.61	5.99
150	3.43	4.39			

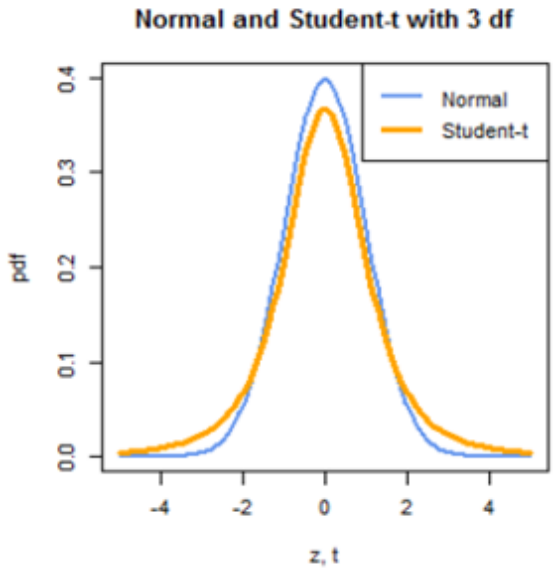
Example: Bowman-Shelton Test for Normality

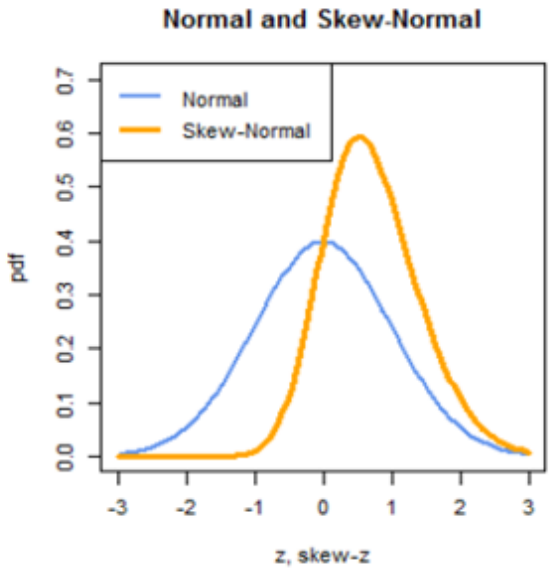
- The average daily temperature has been recorded for 200 randomly selected days, with sample skewness 0.232 and kurtosis 3.319
- Test the null hypothesis that the true distribution is normal

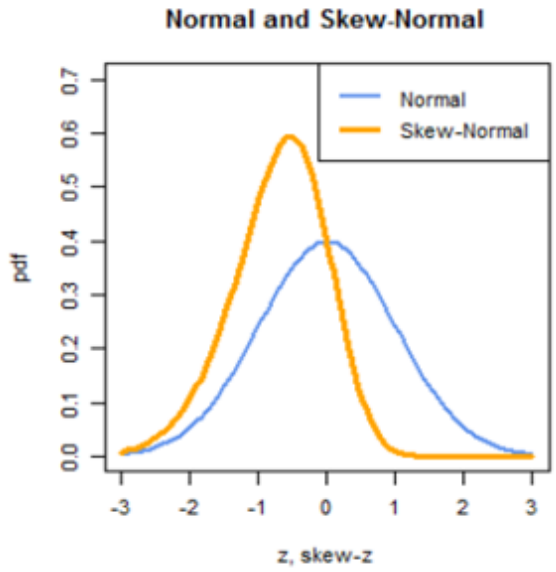
$$B = n\left[\frac{(\text{Skewness})^2}{6} + \frac{(\text{Kurtosis} - 3)^2}{24}\right] = 200\left[\frac{(0.232)^2}{6} + \frac{(3.319 - 3)^2}{24}\right] = 2.642$$

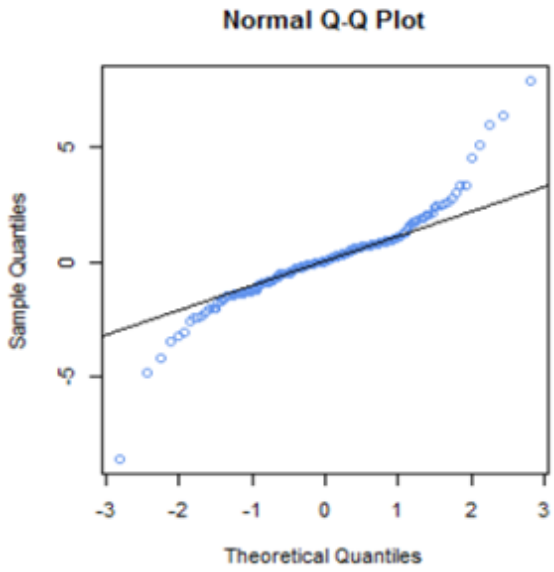
- From Table 16.7 the 10% critical value for n = 200 is 3.48, so there is not sufficient evidence to reject that the population is normal

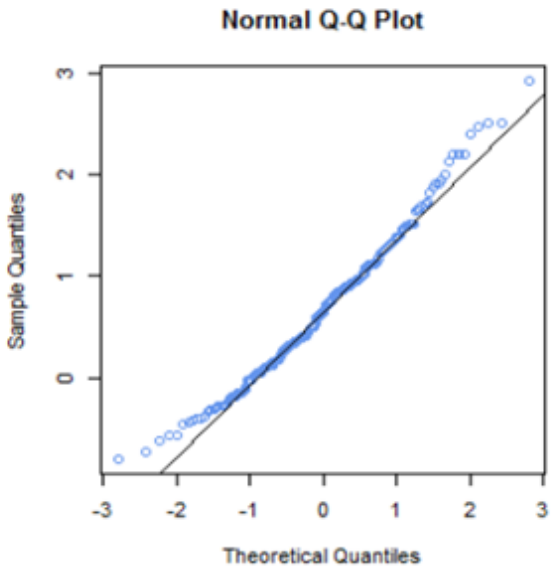
QQ-plot

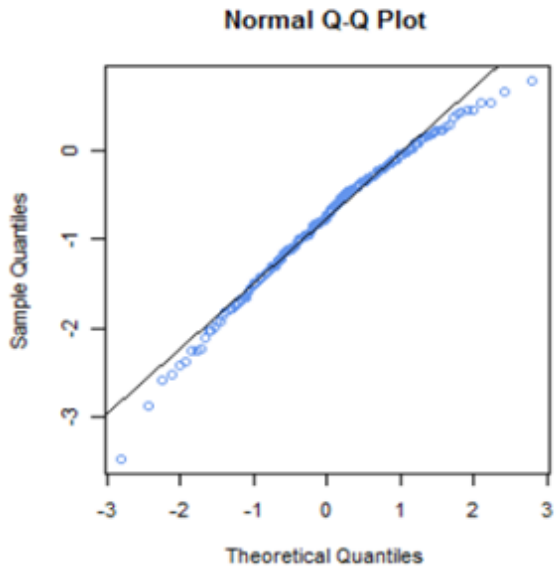

Normal and Student-t with 3 df
Normal
Student-t
pdf
z, t


Normal and Skew-Normal
Normal
Skew-Normal
pdf
z, skew-z


Normal and Skew-Normal
Normal
Skew-Normal
pdf
z, skew-z


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Testing normality: QQ plots (2)

- The subtraction of 0.5 from j is referred to as the continuity correction.

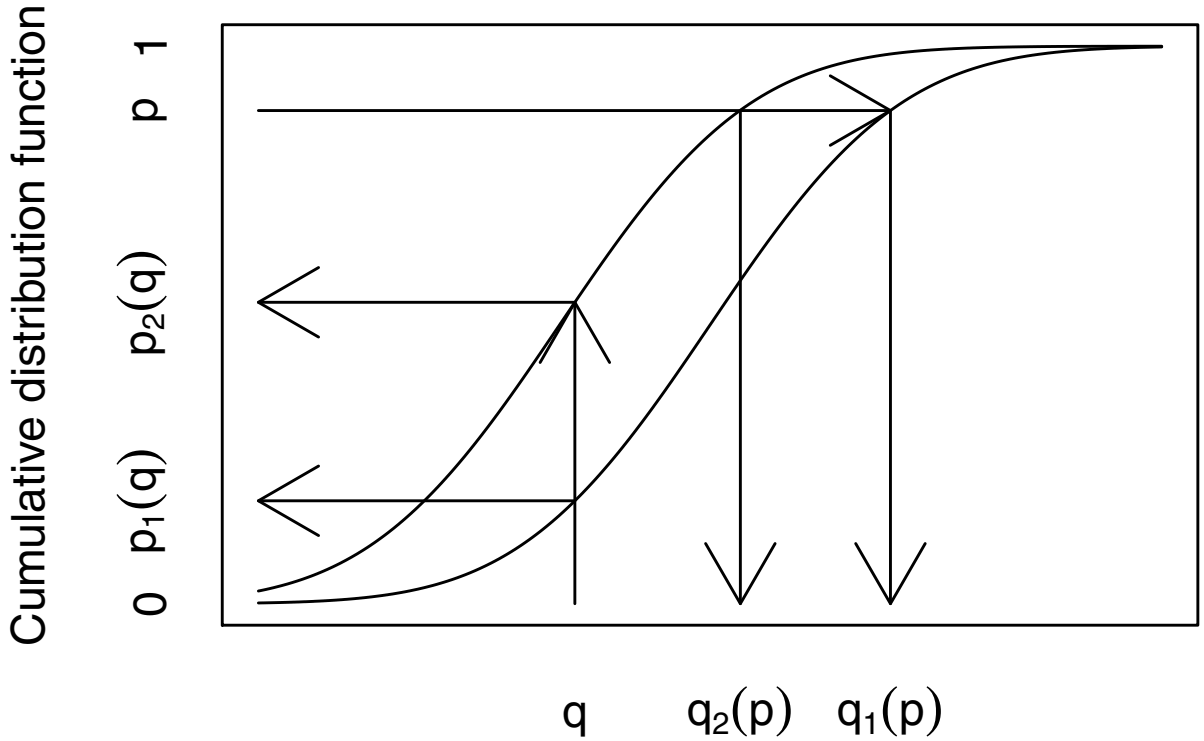
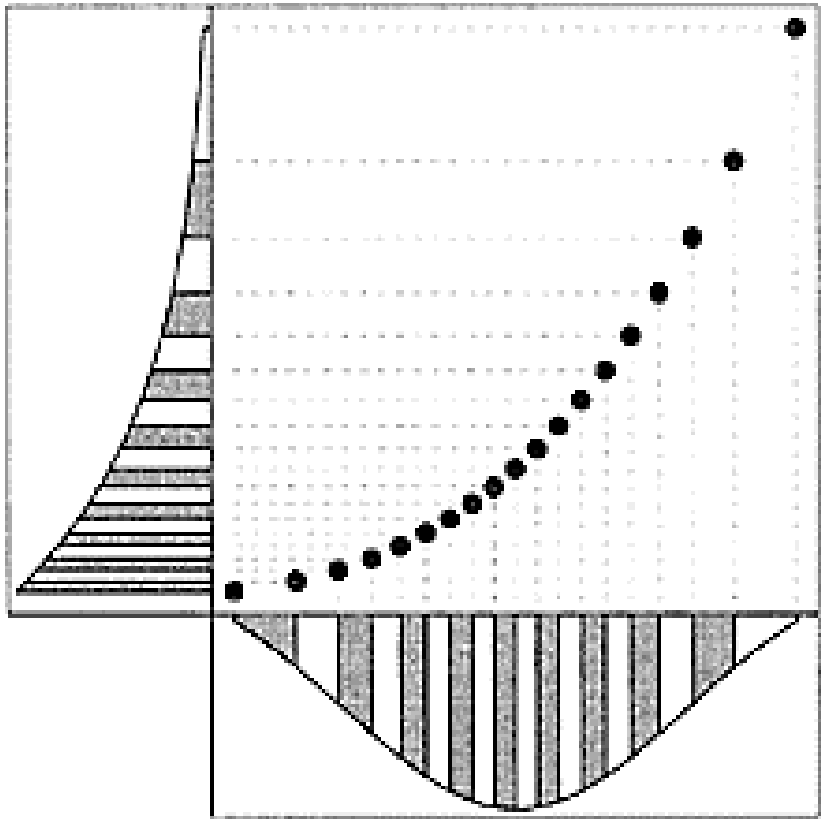


Figure: Quantiles $q_1(p)$, $q_2(p)$ coming from two different distributions

Testing normality: QQ plots (3)

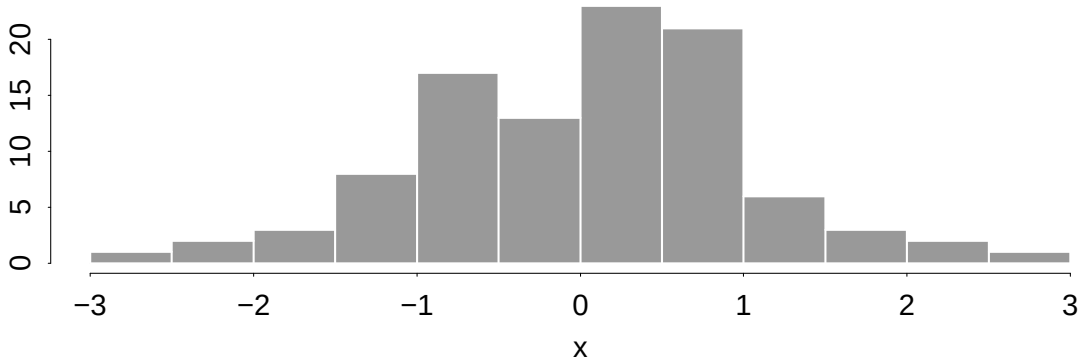
- A quantile-quantile plot (q-q plot) plots the quantiles of one distribution against the quantiles of the other as points
- If the distributions have similar shapes, the points will fall roughly along a straight line.
- If the distributions are different, the points will not lie near a line revealing the domain(s) where the distributions differ.

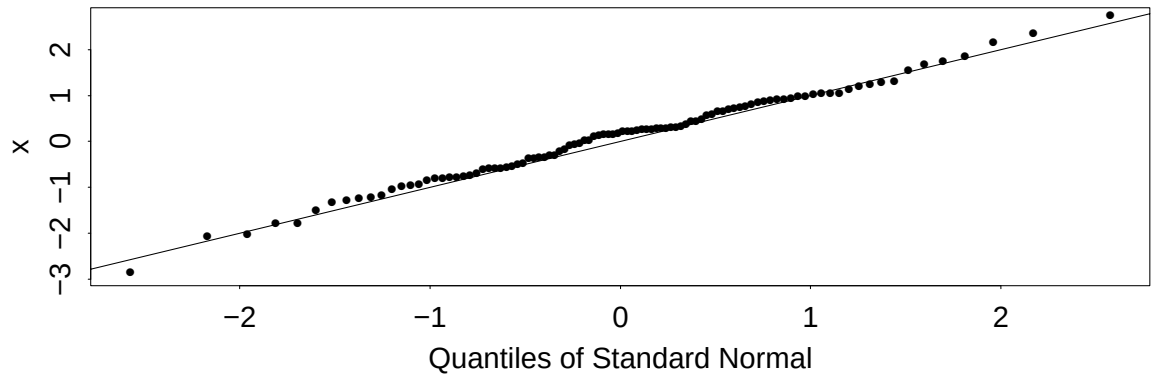


Testing normality: QQ plots (4)

- The subtraction of 0.5 from j is referred to as the continuity correction.
- The following R commands create the normal QQ-plot:

```
> x=rnorm(100)
> par(mfcol=c(2,1))
> hist(x)
> qqnorm(x) # built-in
function
> prob=(c(1:100)-0.5)/100
> z=qnorm(prob)
> y=sort(x)
> cor(y,z)
```





$$\text{cor(y,z)} = 0.9961$$

Testing normality: QQ plots (5)

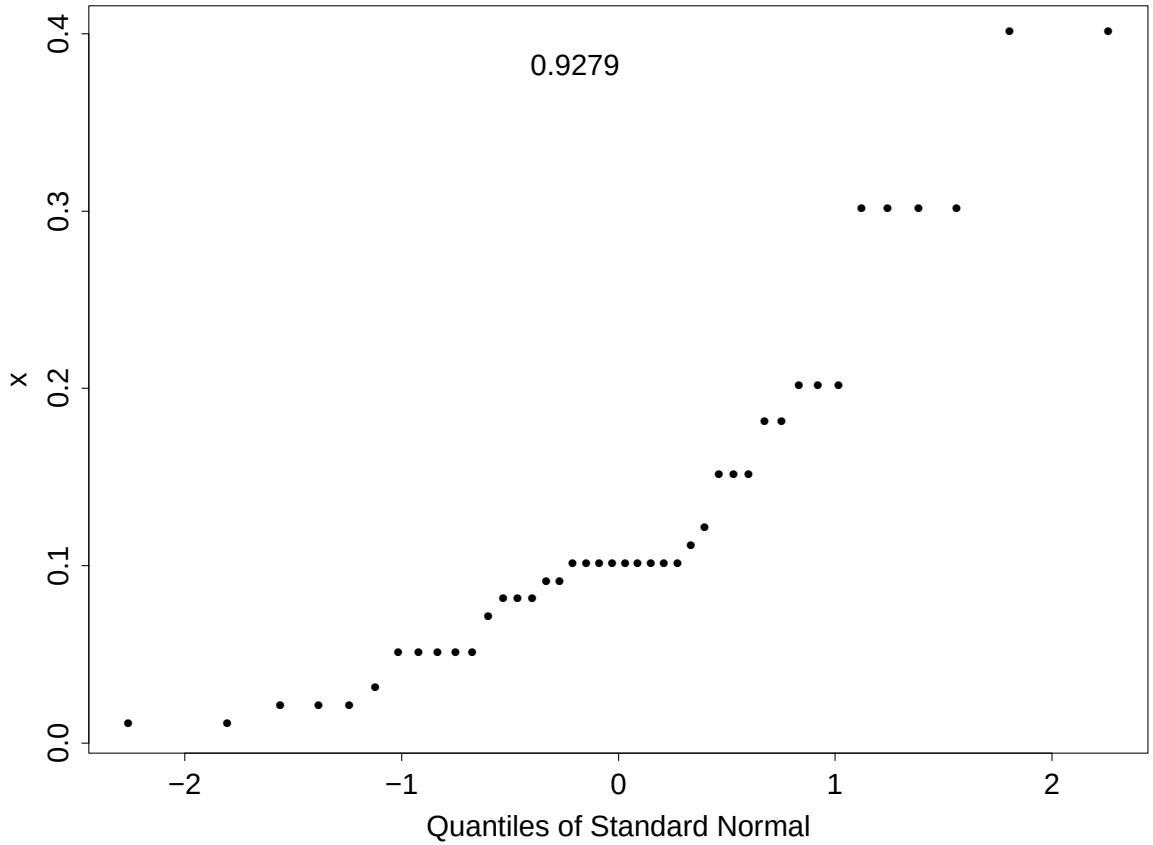
- A simple statistic to check the straight line in a QQ-plot is the correlation coefficient r_Q between $(q_{(j)}, x_{(j)})$
- A table of critical values for r_Q:

Table 4.2 Critical Points for the Q–Q Plot Correlation Coefficient Test for Normality

Sample size n	Significance levels α		
	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Testing normality: QQ plots (6)

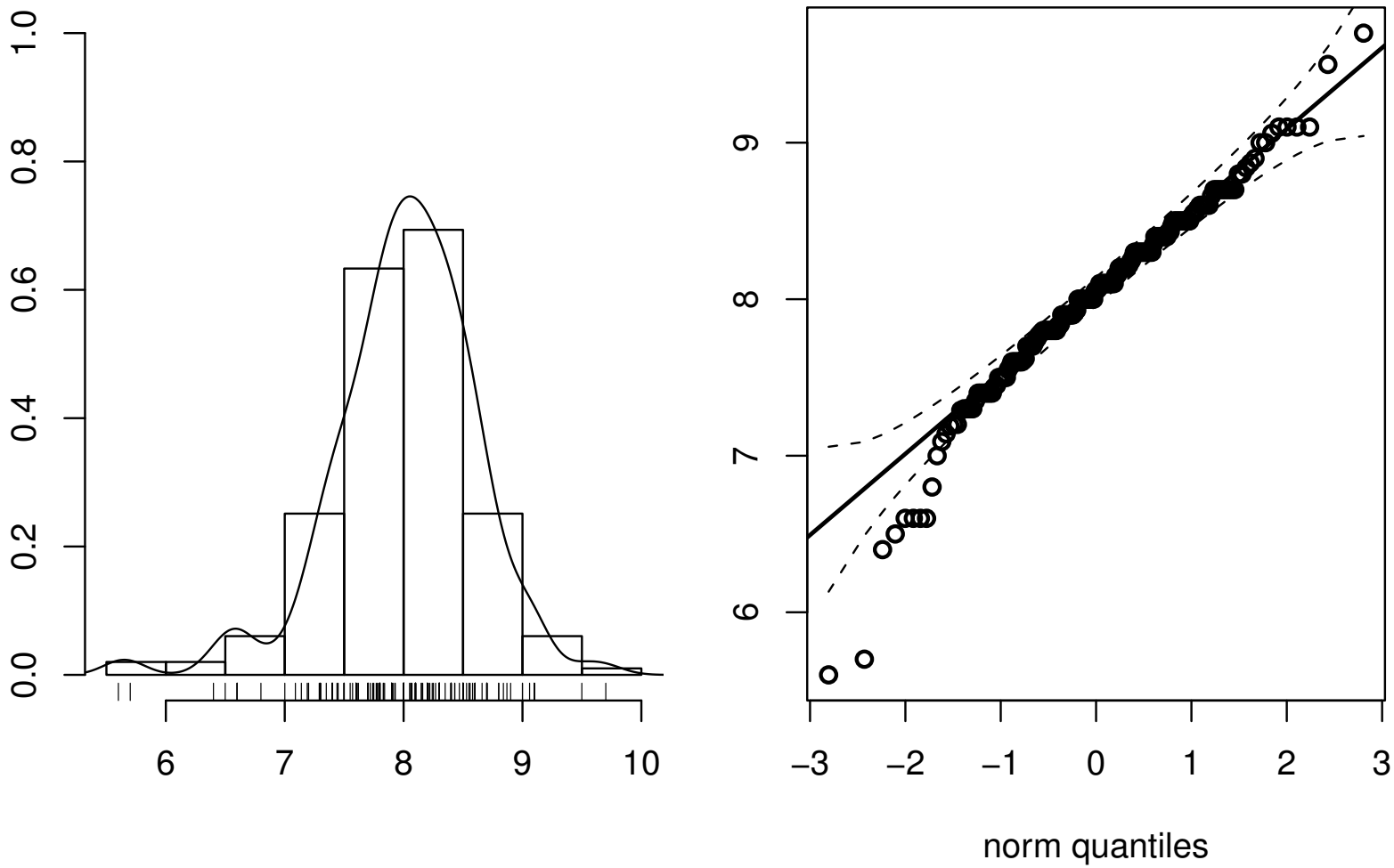
- Another example:



- The correlation coefficient is “too” small (see the Table in Slide 151) => reject normality.

QQ plot with the confidence interval

- Command `qq.plot()` comes from `car` package



Contingency Tables

- Used to classify sample observations according to a pair of attributes
- Also called a cross-classification or cross-tabulation table
- Assume r categories for attribute A and c categories for attribute B
  - Then there are (r x c) possible cross-classifications

$r \times c$ Contingency Table

	Attribute B				
Attribute A	**1**	**2**	**. . .**	**C**	**Totals**
1	O_{11}	O_{12}	...	O_{1c}	R_1
2	O_{21}	O_{22}	...	O_{2c}	R_2
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Totals	C_1	C_2	...	C_c	n

Test for Association

- Consider n observations tabulated in an r x c contingency table
- Denote by O_{ij} the number of observations in the cell that is in the i^{th} row and the j^{th} column
- The null hypothesis is

H_0 : No association exists between the two attributes in the population

- The appropriate test is a chi-square test with (r-1)(c-1) degrees of freedom

Test for Association (2)

- Let R_i and C_j be the row and column totals
- The expected number of observations in cell row i and column j, given that H_0 is true, is

$$E_{ij} = \frac{R_i C_j}{n}$$

- A test of association at a significance level α is based on the chi-square distribution and the following decision rule

$$\text{Reject } H_0 \text{ if } \quad \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)c-1),\alpha}$$

Contingency Table Example

Left-Handed vs. Gender

- Dominant Hand: Left vs. Right
- Gender: Male vs. Female

H_0: There is no association between hand preference and gender

H_1: Hand preference is not independent of gender

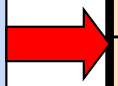
Contingency Table Example (2)

Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed



Gender	Hand Preference		
	Left	Right	
Female	12	108	120
Male	24	156	180
	36	264	300

Logic behind the Test

H_0: There is no association between hand preference and gender

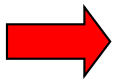
H_1: Hand preference is not independent of gender

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
- The two proportions above should be the same as the proportion of left-handed people overall

Finding Expected Frequencies

120 Females, 12 were left handed

180 Males, 24 were left handed



Overall:

P(Left Handed) = 36/300 = .12

If no association, then

P(Left Handed | Female) = P(Left Handed | Male) = .12

So we would expect 12% of the 120 females and 12% of the 180 males to be left handed…

i.e., we would expect (120)(.12) = 14.4 females to be left handed
(180)(.12) = 21.6 males to be left handed

Expected Cell Frequencies

- Expected cell frequencies:

$$E_{ij} = \frac{R_i C_j}{n} = \frac{(i^{th}\ \text{Row total})(j^{th}\ \text{Column total})}{\text{Total sample size}}$$

Example:

$$E_{11} = \frac{(120)(36)}{300} = 14.4$$

Observed vs. Expected Frequencies

Observed frequencies vs. expected frequencies:

Gender	Hand Preference		
	Left	Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \qquad \text{with} \quad \text{d.f.} = (r-1)(c-1)$$

- where:

  O_{ij} = observed frequency in cell (i, j)

  E_{ij} = expected frequency in cell (i, j)

  r = number of rows

  c = number of columns

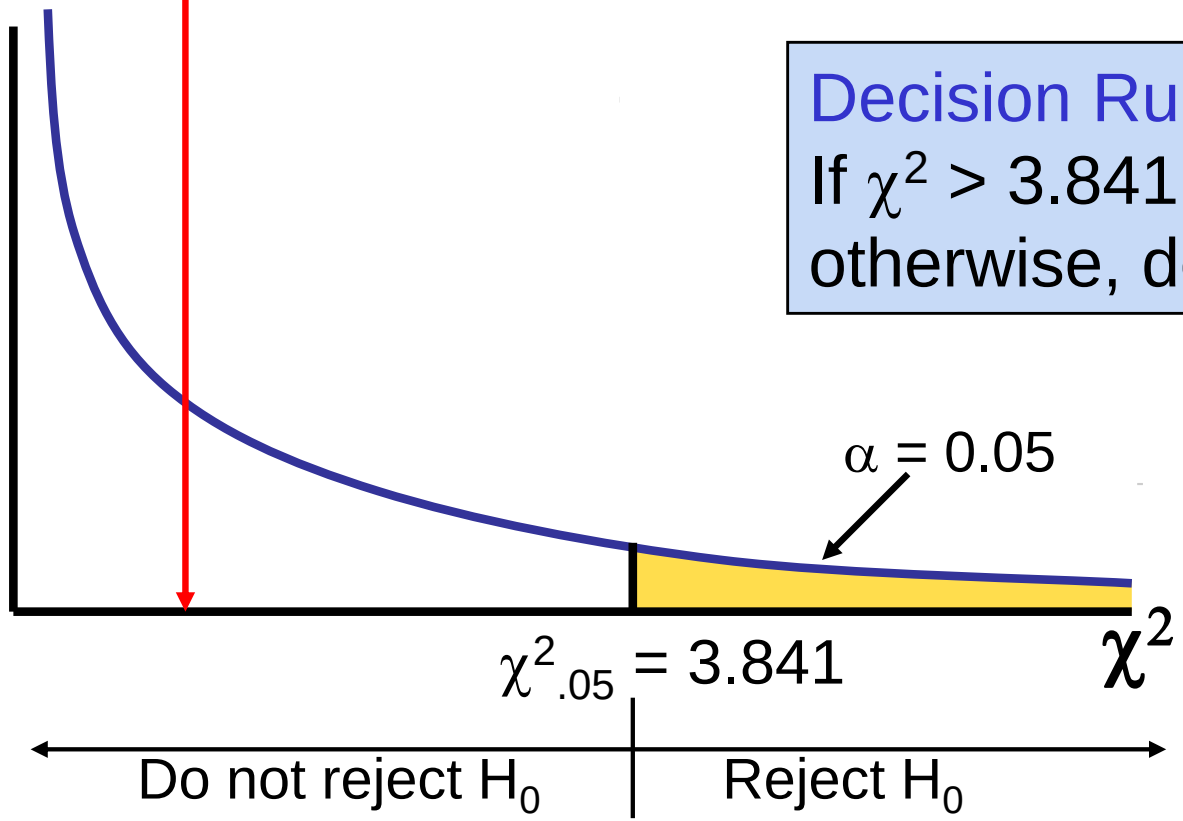
Observed vs. Expected Frequencies

Gender	Hand Preference: Left	Hand Preference: Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

$$\chi^2 = \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.6848$$

Contingency Analysis

$\chi^2 = 0.6848$ with $\text{d.f.} = (r-1)(c-1) = (1)(1) = 1$



Decision Rule:
If $\chi^2 > 3.841$, reject H_0, otherwise, do not reject H_0

Here, $\chi^2 = 0.6848 < 3.841$, so we do not reject H_0 and conclude that gender and hand preference are not associated

Test for Association in R

```
> seatbelt = rbind(c(56,8),c(2,16))
> seatbelt
     [,1] [,2]
[1,]   56    8
[2,]    2   16
> chisq.test(seatbelt)
Pearson's Chi-squared test with Yates' continuity correction
data:  seatbelt
X-squared = 36.00, df = 1, p-value = 1.978e-09
```

Acknowledgments

Slides are adapted from those accompanying Newbold's textbook

References

See Chapters 1-8 of [**?**] for more.