ESTIMATORS

Statistical Inference

- Methods used to infer things about the population.

Statistical Significance

- We use tools of statistical inference to determine if the results are **statistically significant**.

- Note that statistical significance *does not necessarily imply practical significance.*

Tools for Statistical Inference

- Confidence intervals
- Hypothesis/Significance Tests

Estimators-What are they?

$\hat{\boldsymbol{\theta}}$, a statistic value obtained from a sample, is called an **estimator** for the corresponding population parameter $\boldsymbol{\theta}$.

Example: $\overline{X}$, sample mean, estimates, μ,the population mean.

An **unbiased** estimator of a parameter $\boldsymbol{\theta}$:

- This means that the center of the sampling distribution of $\hat{\theta}$, or the average of all $\hat{\theta}$ values, corresponds to the population parameter value, θ.

A **consistent** estimator of a parameter θ:

- This means that as we increase the sample size, n, the value based on a sample, $\hat{\theta}$, gets closer and closer to the value of the corresponding population parameter value, θ.

Example: Show that $\bar{X}$is an unbiased estimator of μ. (So, we need to show that $E(\bar{X}) = \mu$.)

Example: $\bar{X}$is also a consistent estimator of μ.

So, this means that $\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \varepsilon) = 0$ for all $\varepsilon > 0$.

We will initially discuss these population parameters, their sample estimators, and the sampling distributions of their estimators:

Measure	**Population Parameter**	**Sample Statistic (Estimator)**
Mean	μ	$\bar{X}$
Variance	σ^2	s^2

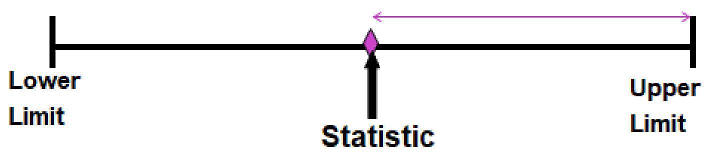
CONFIDENCE INTERVALS: GENERAL IDEA

- The range of values given by a confidence interval:
  - Considers the variation sample **statistics** may have from one sample to another (**the sampling error**)
  - Is based on observations from **one** sample
  - Lets us know how close this particular sample **statistic (estimator)** value may be to **the** unknown population **parameter** value(**margin of sampling error**)
  - Is stated in terms of a **probability** (**confidence level**)

Such as 95% confident, 99% confident, etc. A symmetric confidence interval

<u>Symmetric Confidence Intervals</u>

- The ***statistic (estimator)*** will be the center of a symmetric **confidence interval.**



<u>The Confidence Level</u>

- $a \le \theta \le b$, or $[a, b]$

  - Remember that θ has a constant, usually unknown, value and is not random.
  - The interval is based on data from a random sample and is therefore random.

- $C = 1 - \alpha$

  - That is $\mathrm{P}(\theta \epsilon [a, b]) = 1 - \alpha = C$.
  - The value α represents the probability that the confidence interval **does not cover** θ.

The general formula for all **<u>symmetric</u>** confidence intervals is:

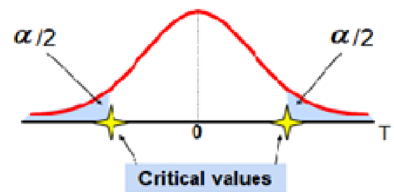
Where:

- The **sample statistic (estimator),** $\boldsymbol{\hat{\theta}}$ **,** corresponds to the population parameter of interest, θ.
- **Critical Value** is a standardized value based on the shape of the sampling distribution, such as $t(\upsilon)$, and the desired confidence level, $1 - \alpha$.
- $\boldsymbol{SD(\hat{\theta})}$ **, the Standard Error** is the standard deviation of the estimator (or an estimation of it).

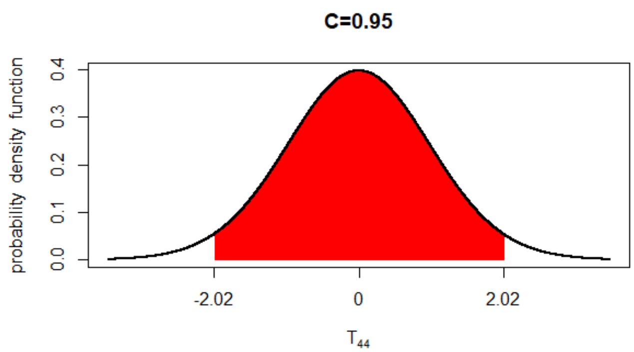
CONFIDENCE INTERVALS FOR A MEAN

The lower and upper limits of the confidence interval for the mean are found by:

$$\overline{X} \pm t(^{\alpha}/_{2}; n-1)\frac{s}{\sqrt{n}}$$



We refer to this interval as a $(1 - \alpha) \times 100\%$ confidence interval for parameter μ.

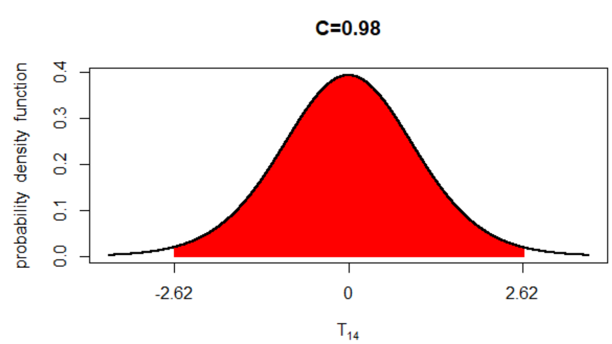


Critical values based on $t(44)$and 95% confidence level:

Example: A tobacco company claims that its best-selling cigarettes contain at most 40 mg of nicotine. Researchers randomly select 15 of these cigarettes and test the nicotine content. The mean is 42.6 mg with standard deviation 3.7 mg. Previous evidence indicates that nicotine content is normally distributed.

a) Find a 98% confidence interval for the true average nicotine level in the company's best-selling cigarettes.

The upper and lower limits of the 98% confidence interval for μ are:



b) Is there evidence to dispute the company's claim, i.e. can we provide statistically significant evidence that there is actually more than 40 mg of nicotine per cigarette, on average?

How do we answer this type of question? **A hypothesis test!**

HYPOTHESIS TESTS

The Set of Hypotheses Used in a Hypothesis Test: (H_0 and H_a)

- A hypothesis is a claim (assertion) about a **population parameter**.
- Construct a set of opposing hypotheses:
  - **The Null Hypothesis(H_0)**
  - **The Alternative Hypothesis (H_A)**

Hypotheses: GENERAL idea

Let $\boldsymbol{\theta_0}$ be a specific number (parameter value) that we are testing.

- Null Hypothesis

$$\boldsymbol{H_0 : \theta = \theta_0}$$

- Alternative Hypothesis (may be one or two sided)

One-sided

$\boldsymbol{H_A : \theta < \theta_0}$ $\boldsymbol{H_A : \theta > \theta_0}$

(Lower-Tail) OR (Upper-Tail) OR

Two-sided

$$\boldsymbol{H_A : \theta \neq \theta_0}$$

Connection Between Hypotheses and Statistical Significance

- When we are **able to reject H_0**

- When our evidence is **not strong enough to reject H_0**

- This decision is associated with a specific **probability**.
- How do we decide whether or not to reject H_0?
  - Begin with the assumption that the null hypothesis is **true** (<u>similar to the notion of innocent until proven guilty</u>).
  - Use a hypothesis test (or confidence interval)!

THE HYPOTHESIS TESTING PROCESS

- **Hypotheses:** Develop a set of hypotheses that you wish to test.
- **Significance Level:** Choose which significance level (α) that you wish to use. (How strong do you want your evidence to be?)
- **Data:** Collect data or refer to given data.
- **Test Statistic:** Find the value of the appropriate test statistic, if needed.
- **Confidence Interval:** Find the lower and upper limits of the confidence interval, if needed.
- **Decision:** Determine if the evidence is strong enough to reject H_0.
  1. Compare Probabilities (Use the P-value)
  2. Use the confidence interval (for two-tailed tests only)
  3. Compare Two Standardized Values (Use a critical value)
- **Conclusion:** Relate your decision back to the original problem.

Test statistic: general idea

- For any hypothesis test we have an assumption that we are working with a random variable whose distribution is based on the sampling distribution of the estimator.
- The **test statistic** is a specific value of the random variable that measures the relative difference between the estimated value from the sample (the sample statistic) and the parameter value being tested in the hypotheses.

Example:

<u>The Test Statistic, P-values and Critical Values</u>

- If the sample statistic is **close** to the stated population parameter,…
- If the sample statistic is **far** from the stated population parameter,…
- How far is "far enough" to reject H_0?
  - The significance level (or associated critical value) of a test creates a cut-off point for decision making -- it answers the question of how far is far enough.

<u>The Significance Level</u>

α=significance level

METHOD 1: COMPARING PROBABILITIES
(Find the p-value and compare it to the significance level)

Decision Rule:

<u>What is the P-value?</u>

- <u>Assuming H_0 is true</u>, the P-value is the probability of obtaining data showing a difference **equal to or larger** than that observed in our **sample** (it is a ***conditional*** *probability*).

- The P-value is also called the **"observed level of significance."**

METHOD 2: SYMMETRIC CONFIDENCE INTERVALS AND TWO-SIDED TESTS

Decision Rule:

TYPE I AND TYPE II ERRORS; POWER OF A TEST

	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
~~Accept~~ H_0 Do not reject	Correct decision	Type II error

$\alpha =$ $\qquad\qquad$ $1 - \alpha =$

$\beta =$ $\qquad\qquad$ $1 - \beta =$

T TESTS FOR MEANS

Assumptions for T tests

- Samples are random and drawn from Normal populations.
  - (In practice, a perfect Normal population is rare. The distribution of the data should show no clear departures form normality-it should be unimodal, roughly symmetric, and contain no outliers.)
  - Population variances are unknown.

P-value for a T test

- Assume that $\mathbf{H_0}$ is true. → This implies that the **center** of the sampling distribution is equal to the value stated in the hypotheses.
- Using a t test assumes that the sampling distribution is a ***t*-distribution**.

Calculate the **T test-statistic** then find the appropriate probability (**P-value**) based on the *t-distribution.*

- For a two-sided test we find the area in both tails.

  $$P - value = \mathbf{2} * \boldsymbol{P}(\boldsymbol{t}(\boldsymbol{v}) > |\boldsymbol{t0}|)$$

- For a one-sided test find the area in the appropriate tail.
  - $P - value = \boldsymbol{P}(\boldsymbol{t}(\boldsymbol{v}) > \boldsymbol{t0})$ for an upper-tail test
  - $P - value = \boldsymbol{P}(\boldsymbol{t}(\boldsymbol{v}) < \boldsymbol{t0})$ lower-tail test

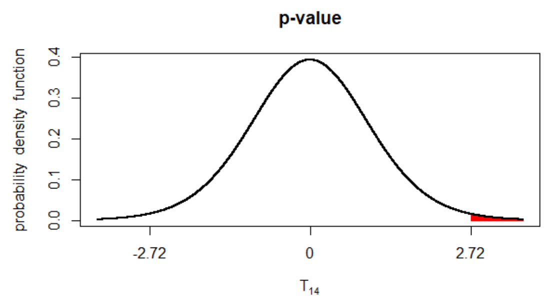
One sample T test for a mean

We are testing the null hypothesis $\boldsymbol{H_0}$:

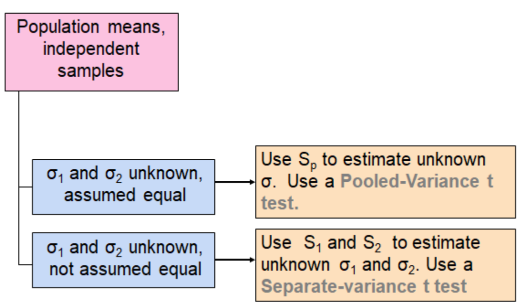
The test-statistic $t_0 \;=$ follows the $t(\upsilon = n - 1)$ distribution.

Example A tobacco company claims that its best-selling cigarettes contain at most 40 mg of nicotine. Researchers randomly select 15 of these cigarettes and test the nicotine content. The mean is 42.6 mg with standard deviation 3.7 mg. Previous evidence indicates that nicotine content is normally distributed.

b) Is there evidence to dispute the company's claim, i.e., can we provide statistically significant evidence that there is actually more than 40 mg of nicotine per cigarette, on average?



Tests for Difference Between Two Means: Independent Samples



Difference Between Two Means

- Independent samples are selected from two normal populations—one with mean μ_1 and variance σ_1^2 and the other with mean μ_2 and variance σ_2^2.
- A random sample of size n_1 with mean $\overline{\boldsymbol{X_1}}$ and variance s_1^2 is drawn from the first population and a random sample of size n_2 with mean $\overline{\boldsymbol{X_2}}$ and variance s_2^2 is drawn independently from the second population.
- The statistic (estimator) for the difference $\mu_1 - \mu_2$ is $\overline{\boldsymbol{X_1}}$- $\overline{\boldsymbol{X_2}}$.
    - The sampling distribution will be a t-distribution.
- We test the null hypothesis H_0: $\mu_1 - \mu_2$ =D.
    - We often use D=0 in the hypotheses (representing no average difference).

Hypothesis tests for μ_1 - μ_2 with $\underline{\sigma 1 \text{ and } \sigma 2}$ unknown and <u>assumed equal</u>

Assumptions:

- Samples are random and independent.
- Populations are Normal or samples are not distinctly non-Normal.
- Population variances are unknown but assumed equal.

The **pooled variance** is:

$$S_p^2 = \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{(n_1 - 1) + (n_2 - 1)}$$

- The **test statistic** is $t_0 = \frac{(\overline{X}_1 - \overline{X}_2) - D}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $\boldsymbol{t_0 \sim t(n_1 + n_2 - 2)}$.

The l**ower** and **upper limits** of the **confidence interval for $\mu_1 - \mu_2$** are determined by:

$$(\overline{X}_1 - \overline{X}_2) \pm t(\alpha/2\,;\, n_1 + n_2 - 2)\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

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:

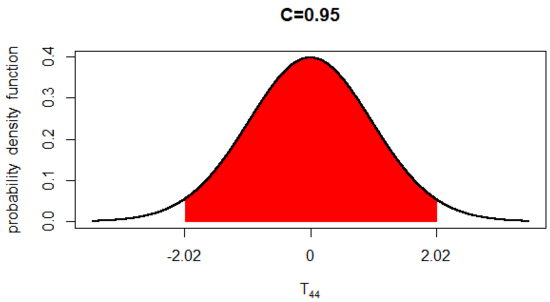
	NYSE	NASDAQ
Sample size	21	25
Sample mean	3.27%	2.53%
Sample std dev	1.30 %	1.16%

Assuming both populations are approximately normal with equal variances, is there a difference in mean yield (α = 0.05)?

$$t_0 = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(3.27 - 2.53) - 0}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = 2.040$$

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

95% Confidence Interval for $\mu_{NYSE} - \mu_{NASDAQ}$: $\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} =$



H**ypothesis tests for μ_1 - μ_2 with $\sigma 1$ and $\sigma 2$ unknown, not assumed equal**

Assumptions:

- Samples are random and independent.
- Populations are Normal or samples are not distinctly non-Normal.
- Population variances are unknown and are not assumed to be equal

The test statistic $t_0 = \dfrac{(\overline{X}_1 - \overline{X}_2) - D}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$

follows a t-distribution with v (degrees of freedom) determined by the ***Satterthwaite approximation***:

$$\nu = \frac{\left(\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}\right)^2}{\frac{\left(\frac{S_1{}^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2{}^2}{n_2}\right)^2}{n_2-1}}$$

Example (contd): You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? Assuming both populations are approximately normal with unequal variances, is there a difference in mean yield (α = 0.05)?

$$t = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}} = \frac{(3.27-2.53)-0}{\sqrt{\left(\frac{1.30^2}{21}+\frac{1.16^2}{25}\right)}} = 2.019$$

$$\nu = \frac{\left(\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}\right)^2}{\frac{\left(\frac{S_1{}^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2{}^2}{n_2}\right)^2}{n_2-1}} = \frac{\left(\frac{1.30^2}{21}+\frac{1.16^2}{25}\right)^2}{\frac{\left(\frac{1.30^2}{21}\right)^2}{20}+\frac{\left(\frac{1.16^2}{25}\right)^2}{24}} = 40.57$$

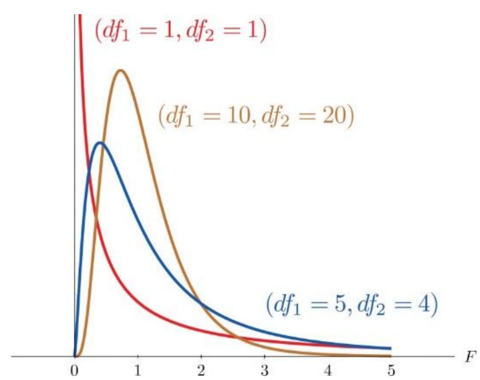
Equal or unequal variances—how do we determine this?

F TESTS FOR COMPARING VARIANCES

Assumptions for comparing two variances

- Independent samples are selected from two normal populations—one with mean μ_1 and variance σ_1^2 and the other with mean μ_2 and variance σ_2^2.
- A random sample of size n_1 with variance s_1^2 is drawn from the first population and a random sample of size n_2 with variance s_2^2 is drawn independently from the second population.
  - (Note: The distribution of $\frac{(n_1-1)s_1^2}{\sigma_1^2}$ follows a χ^2distribution with $n_1 - 1$ degrees of freedom and the sampling distribution$\frac{(n_2-1)s_2^2}{\sigma_2^2}$follows a χ^2distribution with $n_2 - 1$ degrees of freedom.)
- The ratio of these variances ${s_1^2}/{s_2^2}$ is a biased estimator (overestimate) of the parameter ${\sigma_1^2}/{\sigma_2^2}$.
- $\mathrm{F} = \left({s_1^2}/{\sigma_1^2}\right)\Big/\left({s_2^2}/{\sigma_2^2}\right)$ follows an F distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

What is the F distribution?



The F test for comparing variances

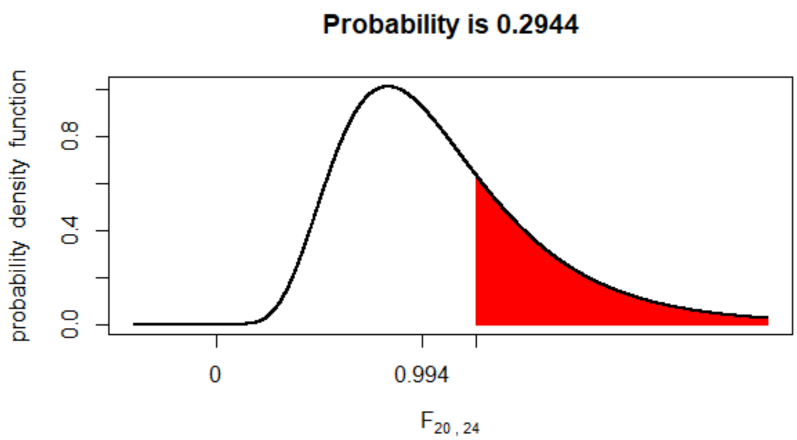
- The test statistic for testing $\boldsymbol{H_0}: \boldsymbol{\sigma_1^2 = \sigma_2^2}$ $(\boldsymbol{or\ \sigma_1^2/\sigma_2^2 = 1})$ is $\boldsymbol{F_0 = \frac{s_1^2}{s_2^2}}$.
- $F_0 = \frac{s_1^2}{s_2^2}$ follows an $\boldsymbol{F(n_1 - 1, n_2 - 1)}$ distribution.

Example:

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

Assuming both populations are approximately normal, should we assume that the variances are **equal or unequal**?

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{1.3^2}{1.16^2} = 1.256$$



CONCLUSION: