Statistics for Business

Sampling Distributions, Interval Estimation and Hypothesis Tests.

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: July 15, 2015. **This Draft**: August 28, 2023.

Lecture Outline

- Simple random sampling
- Distribution of the sample average
- Large sample approximation to the distribution of the sample mean
  - Law of Large Numbers
  - Central Limit Theorem
- Estimation of the population mean
  - Unbiasedness
  - Consistency
  - Efficiency
- Hypothesis test concerning the population mean
- Confidence intervals for the population mean
  - Using the t-statistic when n is small
- Comparing means from different populations

Sampling

- A ***population*** is a collection of all the elements of interest, while a ***sample*** is a subset of the population.
- The reason we select a sample is to collect data to answer a research question about a population.
- The sample results provide only **estimates** of the values of the population characteristics. With *proper sampling methods*, the sample results can provide "good" estimates of the population characteristics.
- A ***random sample*** from an infinite population is a sample selected such that the following conditions are satisfied:
  - Each element selected comes from the population of interest.
  - Each element is selected *independently*.
  - ★ If the population is finite, then we sample with replacement...

Simple Random Sampling – I

- ***Simple random sampling*** means that n objects are drawn randomly from a population and each object is equally likely to be drawn
- Let $Y_1, Y_2, ..., Y_n$ denote the 1st to the n th randomly drawn object. Under simple random sampling
  - The marginal probability distribution of Y_i is the same for all $i = 1, 2, ..., n$ and equals the population distribution of Y.
  - ⋆ because $Y_1, Y_2, ..., Y_n$ are drawn randomly from the **same** population.
  - Y_1 is distributed independently from $Y_2, ..., Y_n$. knowing the value of Y_i does not provide information on Y_j for $i \neq j$
- When $Y_1, Y_2, ..., Y_n$ are drawn from the same population and are independently distributed, they are said to be ***I.I.D. random variables***

Simple Random Sampling – II

Example

- Let G be the gender of an individual ($G = 1$ if female, $G = 0$ if male)
- G is a Bernoulli r.v. with $\mathrm{E}(G) = \mu_G = \Pr(G = 1) = 0.5$
- Suppose we take the population register and randomly draw a sample of size n
  - The probability distribution of G_i is a Bernoulli with mean 0.5
  - G_1 is distributed independently from $G_2, ..., G_n$
- Suppose we draw a random sample of individuals entering the building of the accounting department
  - This is not a sample obtained by simple random sampling and $G_1, G_2, ..., G_n$ are not i.i.d
  - Men are more likely to enter the building of the accounting department!

The Sampling Distribution of the Sample Average – I

- The ***sample average*** $\bar{Y}$ of a randomly drawn sample is a random variable with a probability distribution called the ***sampling distribution***

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  - The individuals in the sample are drawn at random.
  - Thus the values of $(Y_1, Y_2, \cdots, Y_n)$ are random
  - Thus functions of $(Y_1, Y_2, \cdots, Y_n)$, such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value
  - The distribution of over different possible samples of size n is called the ***sampling distribution*** of $\bar{Y}$.
  - The mean and variance of are the mean and variance of its sampling distribution, $\mathrm{E}(\bar{Y})$ and $\mathrm{Var}(\bar{Y})$.
  - The concept of the sampling distribution underpins all of statistics/econometrics.

The Sampling Distribution of the Sample Average – II

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Suppose that $Y_1, Y_2, ..., Y_n$ are *I.I.D.* and the mean & variance of the population distribution of Y are respectively μ_Y and σ_Y^2
  - The mean of (the sampling distribution of) $\bar{Y}$ is

$$\mathrm{E}(\bar{Y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i) = \frac{1}{n}n\mathrm{E}(Y) = \mu_Y$$

  - The variance of (the sampling distribution of) $\bar{Y}$ is

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) + 2\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathrm{Cov}(Y_i, Y_j) \\ &= \frac{1}{n^2}n\mathrm{Var}(Y) + 0 = \frac{1}{n}\mathrm{Var}(Y) = \frac{\sigma_Y^2}{n}\end{aligned}$$

The Sampling Distribution of the Sample Average – III

Example

- Let G be the gender of an individual ($G = 1$ if female, $G = 0$ if male)
- The mean of the population distribution of G is

$$\mathrm{E}(G) = \mu_G = \Pr(G = 1) = p = 0.5$$

- The variance of the population distribution of G is

$$\mathrm{Var}(G) = \sigma_G^2 = p(1 - p) = 0.5(1 - 0.5) = 0.25$$

- The mean and variance of the average gender (proportion of women) $\bar{G}$ in a random sample with $n = 10$ are

$$\begin{aligned}\mathrm{E}(\bar{G}) &= \mu_G = 0.5 \\ \mathrm{Var}(\bar{G}) &= \frac{1}{n}\sigma_G^2 = \frac{1}{10}0.25 = 0.025\end{aligned}$$

The Finite-Sample Distribution of the Sample Average

- The ***finite sample distribution*** is the sampling distribution that exactly describes the distribution of $\bar{Y}$ for any sample size n.
- In general the exact sampling distribution of $\bar{Y}$ is complicated and depends on the population distribution of Y.
- A special case is when $Y_1, Y_2, ..., Y_n$ are *IID* draws from the $N(\mu_Y, \sigma_Y^2)$, because in this case

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

The Sampling Distribution of the Average Gender $\bar{G}$

- Suppose G takes on 0 or 1 (a Bernoulli random variable) with the probability distribution

$$\Pr(G = 0) = p = 0.5, \quad \Pr(G = 1) = 1 - p = 0.5$$

- As we discussed above:

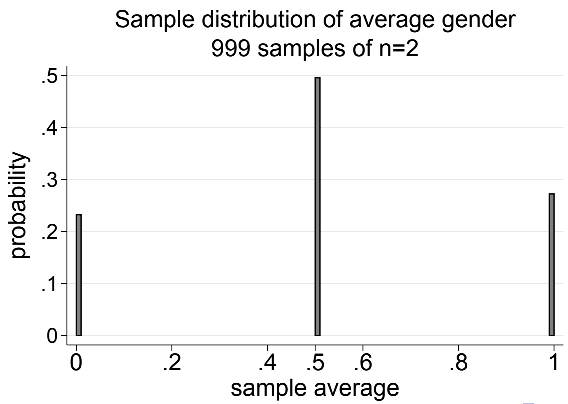
$$\begin{aligned} \mathrm{E}(G) &= \mu_G = \Pr(G = 1) = p = 0.5 \\ \mathrm{Var}(G) &= \sigma_G^2 = p(1 - p) = 0.5(1 - 0.5) = 0.25 \end{aligned}$$

- The sampling distribution of $\bar{G}$ depends on n.
- Consider $n = 2$. The sampling distribution of $\bar{G}$ is
  - $\Pr(\bar{G} = 0) = 0.5^2 = 0.25$
  - $\Pr(\bar{G} = 1/2) = 2 \times 0.5 \times (1 - 0.5) = 0.5$
  - $\Pr(\bar{G} = 1) = (1 - 0.5)^2 = 0.25$

The Finite-Sample Distribution of the Average Gender $\bar{G}$

- Suppose we draw 999 samples of $n = 2$:

Sample 1			Sample 1			Sample 3			$\cdots$	Sample 999		
G_1	G_2	$\bar{G}$	G_1	G_2	$\bar{G}$	G_1	G_2	$\bar{G}$		G_1	G_2	$\bar{G}$
1	0	0.5	1	1	1	0	1	0.5		0	0	0



The Asymptotic Distribution of the Sample Average $\bar{Y}$

- Given that the exact sampling distribution of $\bar{Y}$ is complicated and given that we generally use large samples in statistics/econometrics we will often use an approximation of the sample distribution that relies on the sample being large
- The ***asymptotic distribution*** or ***large-sample distribution*** is the approximate sampling distribution of $\bar{Y}$ if the sample size becomes very large: $n \to \infty$.
- We will use two concepts to approximate the large-sample distribution of the sample average
  - The law of large numbers.
  - The central limit theorem.

The Law of Large Numbers (LLN)

Definition (Law of Large Numbers)

Suppose that

1. $Y_i, i = 1, ..., n$ are independently and identically distributed with $\mathrm{E}(Y_i) = \mu_Y$; and
2. large outliers are unlikely i.e. $\mathrm{Var}(Y_i) = \sigma_Y^2 < +\infty$.

Then $\bar{Y}$ will be near μ_Y with very high probability when n is very large ($n \to \infty$)

$$\bar{Y} \xrightarrow{p} \mu_Y.$$

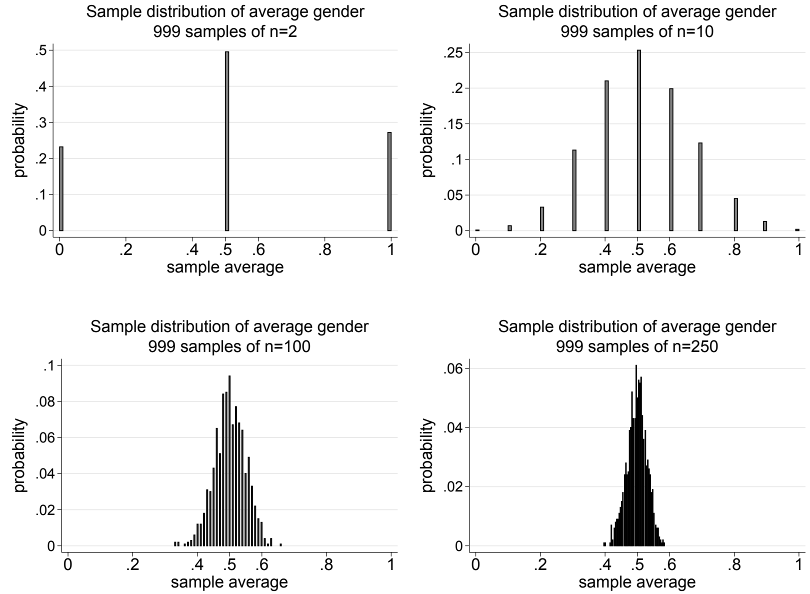
We also say that the sequence of random variables $\{Y_n\}$ converges in probability to the μ_Y, if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr(|\bar{Y}_n - \mu_Y| > \varepsilon) = 0.$$

We also denote this by $\mathrm{plim}(Y_n) = \mu_Y$

The Law of Large Numbers (LLN)

Example: Gender $G \sim Bernoulli(0.5, 0.25)$



The Central Limit Theorem (CLT)

Definition (Central Limit Theorem)

Suppose that

1. $Y_i, i = 1, ..., n$ are independently and identically distributed with $\mathrm{E}(Y_i) = \mu_Y$; and
2. large outliers are unlikely i.e. $\mathrm{Var}(Y_i) = \sigma_Y^2$ with $0 < \sigma_Y^2 < +\infty$.

Then the distribution of the sample average $\bar{Y}$ will be approximately normal as n becomes very large ($n \to \infty$)

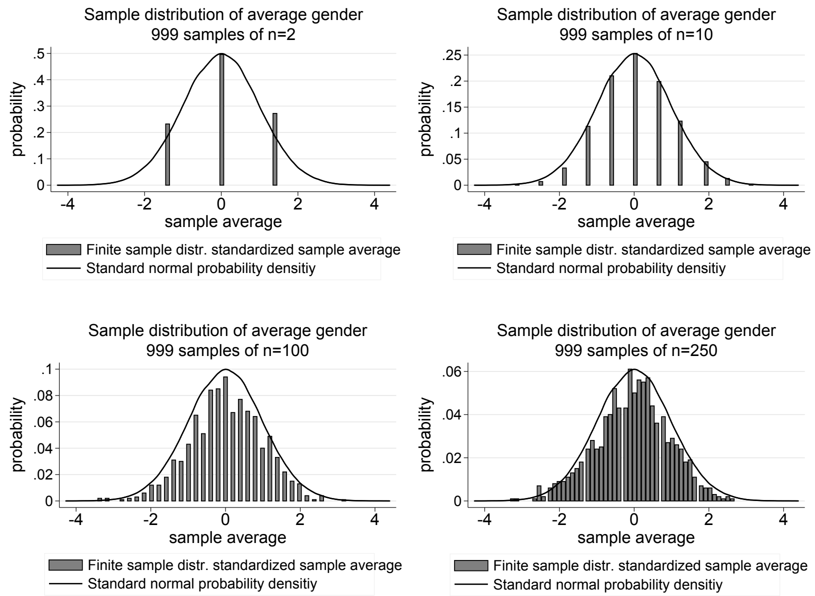
$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right).$$

The distribution of the the standardized sample average is approximately standard normal for $n \to \infty$

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$$

The Central Limit Theorem (CLT)

Example: Gender $G \sim Bernoulli(0.5, 0.25)$



The Central Limit Theorem (CLT)

- How good is the large-sample approximation?

⋆ If $Y_i \sim N(\mu_Y, \sigma_Y^2)$ the approximation is perfect.

⋆ If Y_i is not normally distributed the quality of the approximation depends on how close n is to infinity (how large n is)

⋆ For $n \geq 100$ the normal approximation to the distribution of $\bar{Y}$ is typically very good for a wide variety of population distributions.

Estimators and Estimates

Definition

An ***estimator*** is a function of a sample of data to be drawn randomly from a population.

- An estimator is a random variable because of randomness in drawing the sample. Typically used estimators

$$\text{Sample Average:} \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \ \text{Sample variance: } S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Using a particular sample $y_1, y_2, \ldots, y_n$ we obtain

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

which are ***point estimates***. These are the numerical value of an estimator when it is actually computed using a specific sample.

Estimation of the Population Mean – I

- Suppose we want to know the mean value of Y (μ_Y) in a population, for example
  - The mean wage of college graduates.
  - The mean level of education in Greece.
  - The mean probability of passing the statistics exam.
- Suppose we draw a random sample of size n with $Y_1, Y_2, \ldots, Y_n$ being *IID*
- Possible estimators of μ_Y are:
  - The sample average: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
  - The first observation: Y_1
  - The weighted average: $\tilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \ldots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right)$.
- To determine which of the estimators, $\bar{Y}$, Y_1 or $\tilde{Y}$ is the best estimator of μ_Y we consider 3 properties.
- Let $\hat{\mu}_Y$ be an estimator of the population mean μ_Y

Estimation of the Population Mean – II

1. **Unbiasedness**: The mean of the sampling distribution of $\hat{\mu}_Y$ equals μ_Y
$$\mathrm{E}(\hat{\mu}_Y) = \mu_Y.$$
2. **Consistency**: The probability that $\hat{\mu}_Y$ is within a very small interval of μ_Y approaches 1 if $n \to \infty$
$$\hat{\mu}_Y \xrightarrow{p} \mu_Y \text{ or } \Pr(|\hat{\mu}_Y - \mu_Y| < \varepsilon) = 1$$
3. **Efficiency**: If the variance of the sampling distribution of $\hat{\mu}_Y$ is smaller than that of some other estimator $\tilde{\mu}_Y$, $\hat{\mu}_Y$ is more efficient
$$\mathrm{Var}(\hat{\mu}_Y) \leq \mathrm{Var}(\tilde{\mu}_Y)$$

Estimating Mean Wages – I

- Suppose we are interested in the mean wages (pre tax) μ_W of individuals with a Ph.D. in economics/finance in Europe (true mean $\mu_w = 60K$). We draw the following sample ($n = 10$) by simple random sampling

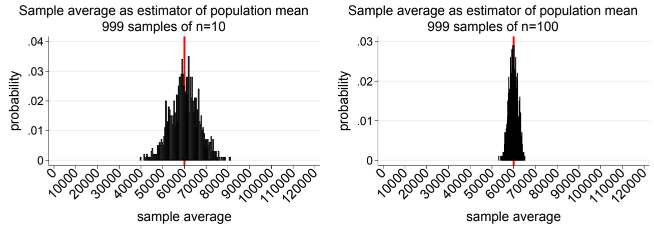
i	1	2	3	4	5
W_i	47281.92	70781.94	55174.46	49096.05	67424.82

i	6	7	8	9	10
W_i	39252.85	78815.33	46750.78	46587.89	25015.71

- The 3 estimators give the following estimates:
  - $\bar{W} = \frac{1}{10}\sum_{i=1}^{10} W_i = 52618.18$
  - $W_1 = 47281.92$
  - $\tilde{W} = \frac{1}{10}\left(\frac{1}{2}W_1 + \frac{3}{2}W_2 + ... + \frac{1}{2}W_9 + \frac{3}{2}W_{10}\right) = 49398.82$
- **Unbiasedness**: All 3 proposed estimators are unbiased

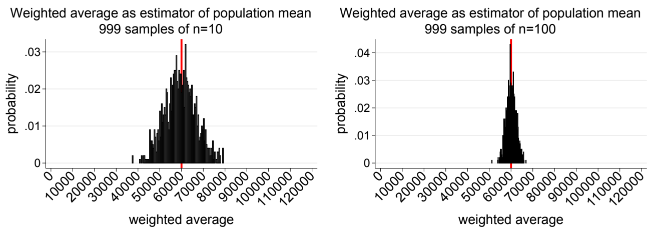
Estimating Mean Wages – II

- **Consistency**:
  - By the law of large numbers $\bar{W} \xrightarrow{p} \mu_W$ which implies that the probability that $\bar{W}$ is within a very small interval of μ_W approaches 1 if $n \to \infty$

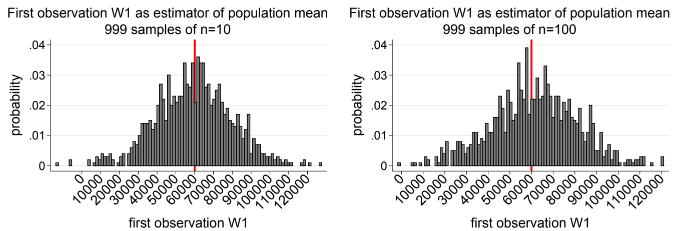


Estimating Mean Wages – III

- $\tilde{W} = \frac{1}{n}\left(\frac{1}{2}W_1 + \frac{3}{2}W_2 + ... + \frac{1}{2}W_{n-1} + \frac{3}{2}W_n\right)$ can also be shown to be consistent



- However W_1 is not a consistent estimator of μ_W.



Estimating Mean Wages – IV

- **Efficiency**: We have that
  - $\text{Var}(\bar{W}) = \frac{1}{n}\sigma_W^2$
  - $\text{Var}(W_1) = \sigma_W^2$
  - $\text{Var}(\tilde{W}) = 1.25\frac{1}{n}\sigma_W^2$
  - So for any $n \geq 2$, $\bar{W}$ is more efficient than W_1 and $\tilde{W}$.
- In fact **$\bar{Y}$ is the Best Linear Unbiased Estimator (BLUE)**: it is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, Y_2, ..., Y_n$

⋆ Let $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} \alpha_i Y_i$ be an unbiased estimator of μ_Y with α_i nonrandom constants. Then $\bar{Y}$ is more efficient than $\hat{\mu}_Y$

$$\text{Var}(\bar{Y}) \leq \text{Var}(\hat{\mu}_Y)$$

Hypothesis Tests

Consider the following questions:

- Is the mean monthly wage of Ph.D. graduates equal to 60000 euros?
- Is the mean level of education in Greece equal to 12 years?
- Is the mean probability of passing the stats exam equal to 1?

These questions involve the population mean taking on a specific value $\mu_{Y,0}$. Answering these questions implies using data to compare a ***null hypothesis*** (a tentative assumption about the population mean parameter)

$$H_0 : \mathrm{E}(Y) = \mu_{Y,0}$$

to an ***alternative hypothesis*** (the opposite of what is stated in the H_0)

$$H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$$

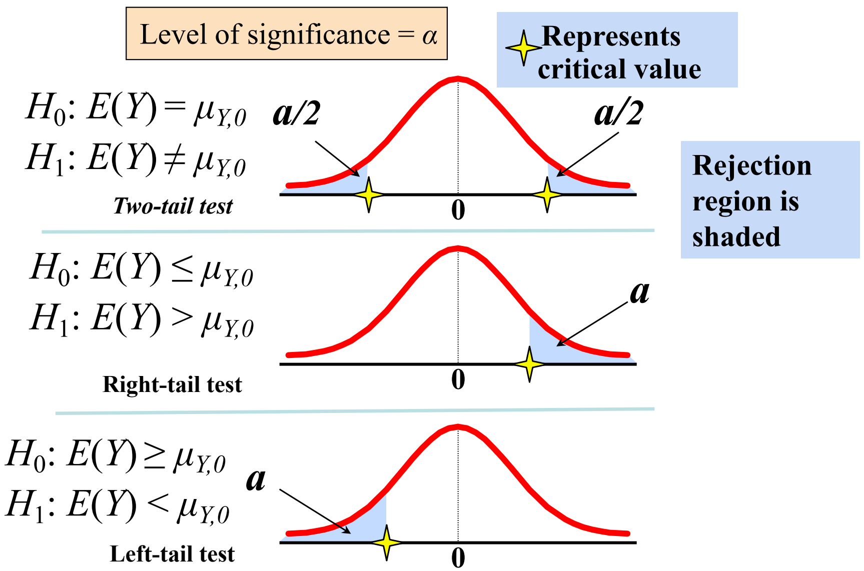
- Alternative Hypothesis as a Research Hypothesis
  - *Example*: A new sales force bonus plan is developed in an attempt to increase sales.
  - **Alternative Hypothesis**: The new bonus plan increase sales.
  - **Null Hypothesis**: The new bonus plan does not increase sales.

Hypothesis Tests: Terminology

- The **hypothesis testing problem** (for the mean): make a provisional decision, based on the evidence at hand, whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test
  - $H_0 : \mathrm{E}(Y) \leq \mu_{Y,0}$ vs. $H_1 : \mathrm{E}(Y) > \mu_{Y,0}$ (1-sided, $>$)
  - $H_0 : \mathrm{E}(Y) \geq \mu_{Y,0}$ vs. $H_1 : \mathrm{E}(Y) < \mu_{Y,0}$ (1-sided, $<$)
  - $H_0 : \mathrm{E}(Y) = \mu_{Y,0}$ vs. $H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$ (2-sided)
- p-value = probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.
- The **significance level** of a test (α) is a pre-specified probability of incorrectly rejecting the null, when the null is true. Typical values are 0.01 (1%), 0.05 (5%), or 0.10 (10%).
  - It is selected by the researcher at the beginning, and determines the ***critical value(s)*** of the test.
  - If the test-statistic falls outside the non-rejection region, we reject H_0.

Hypothesis Tests

The Testing Process and Rejections



Hypothesis Testing using p-values

- The p-value is the probability, computed using the test statistic, that measures the support (or lack of support) provided by the sample for the null hypothesis
  - If the p-value is less than or equal to the level of significance α, the value of the test statistic is in the rejection region.
  - Reject H_0 if the p-value $< \alpha$.
  - See also Annex
- **Rules of thumb**
  - If p-value is less than .01, there is overwhelming evidence to conclude H_0 is false.
  - If p-value is between .01 and .05, there is strong evidence to conclude H_0 is false.
  - If p-value is between .05 and .10, there is weak evidence to conclude H_0 is false.
  - If p-value is greater than .10, there is insufficient evidence to conclude H_0 is false.

Hypothesis Test for the Mean with σ_Y^2 **known** – I

Decision Rules

- The test statistic employed is obtained by converting the sample result ($\bar{y}$) to a z-value

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$$

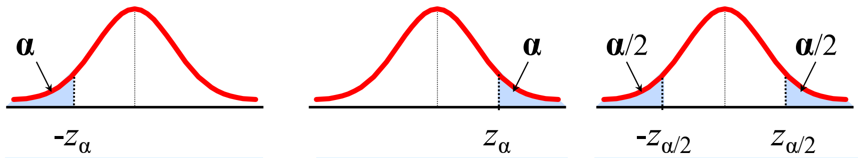
$H_0 : \mathrm{E}(Y) \geq \mu_{Y,0}$ $H_1 : \mathrm{E}(Y) < \mu_{Y,0}$	$H_0 : \mathrm{E}(Y) \leq \mu_{Y,0}$ $H_1 : \mathrm{E}(Y) > \mu_{Y,0}$	$H_0 : \mathrm{E}(Y) = \mu_{Y,0}$ $H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$
Lower-tail Reject H_0 if $z < z_\alpha$	Upper-tail Reject H_0 if $z > z_\alpha$	Two-tailed Reject H_0 if $z < -z_{\alpha/2}$ or if $z > z_{\alpha/2}$

Hypothesis Test for the Mean with σ_Y^2 **known** – II

Decision Rules

Hypothesis Tests for $E(Y)$ $z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(Y) \geq \mu_0$ H_1: $E(Y) < \mu_0$	H_0: $E(Y) \leq \mu_{Y,0}$ H_1: $E(Y) > \mu_{Y,0}$	H_0: $E(Y) = \mu_{Y,0}$ H_1: $E(Y) \neq \mu_{Y,0}$



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

Hypothesis Test for the Mean (σ^2 **known**) – I

Examples

- **Example 1**. A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over \$52 per month. The company wishes to test this claim. Assume $\sigma = 10\$$ is known and let $\alpha = 0.10$. Suppose a sample of 64 persons is taken, and it is found that the average bill \$53.1.
  - Form the hypothesis to be tested

$$H_0 : \mathrm{E}(Y) \leq 52 \quad \text{the } \textit{mean} \textbf{ is not} \text{ over \$52 per month}$$
$$H_1 : \mathrm{E}(Y) > 52 \quad \text{the } \textit{mean} \textbf{ is} \text{ over \$52 per month}$$

  - For $\alpha = 0.10$, $z_{0.10} = 1.28$, so we would reject H_0 if $z > 1.28$.
  - We have $n = 64$ and $\bar{y} = 53.1$, so the test statistic is

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}} = \frac{53.1 - 52}{10/\sqrt{64}} = 0.88 < z_{0.10} = 1.28$$

  Hence H_0 cannot be rejected.

Hypothesis Test for the Mean (σ^2 **known**) – II

Examples

- **Example 2**. We would like to test the claim that the true mean # of TV sets in EU homes is equal to 3 (assuming $\sigma_Y = 0.8$ known). For this purpose a sample of 100 homes is selected, and the average number of TV sets is 2.84. Test the above hypothesis using $\alpha = 0.05$.
  - Form the hypothesis to be tested

$$H_0 : \mathrm{E}(Y) = 3 \quad \text{the } \textit{mean} \text{ \# } \textbf{is} \text{ 3 TV sets per home}$$
$$H_1 : \mathrm{E}(Y) \neq 3 \quad \text{the } \textit{mean} \textbf{ is not} \text{ 3 TV sets per home}$$

  - For $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$, so we would reject H_0 if $|z| > 1.96$.

Hypothesis Test for the Mean (σ^2 **known**) – III

Examples

- We have $n = 100$ and $\bar{y} = 2.84$, so the test statistic is

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}} = \frac{2.84 - 3}{0.8/\sqrt{100}} = \frac{-0.16}{0.08} = -2 < -z_{0.025} = -1.96$$

or $|z| = 2 > 1.96$, Hence H_0 is rejected. We **conclude** that there is sufficient evidence that the mean number of TVs in EU homes is not equal to 3.

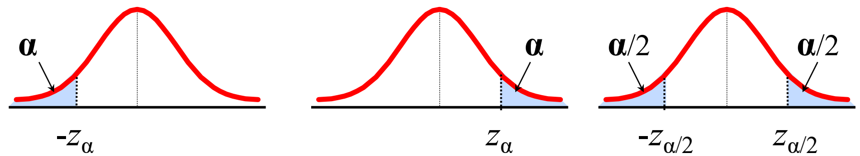
Test for the Mean with σ_Y^2 **unknown** but $n \to \infty$

Decision Rules

- Since $S_Y^2 \xrightarrow{p} \sigma_Y^2$, compute the standard error of $\bar{Y}$, $SE(\bar{Y}) = s_Y/\sqrt{n}$ and construct a t-ratio.

Hypothesis Tests for $E(Y)$ $\quad t = \frac{\bar{Y} - \mu_{Y,0}}{\mathrm{SE}(\bar{Y})} = \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(Y) \geq \mu_0$ H_1: $E(Y) < \mu_0$	H_0: $E(Y) \leq \mu_{Y,0}$ H_1: $E(Y) > \mu_{Y,0}$	H_0: $E(Y) = \mu_{Y,0}$ H_1: $E(Y) \neq \mu_{Y,0}$



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

Test for the Mean with σ_Y^2 **unknown** but $n \to \infty$

Example

- Suppose we would like to test

$$H_0 : \mathrm{E}(W) = 60000, \qquad H_1 : \mathrm{E}(W) \neq 60000,$$

using a sample of 250 individuals with a Ph.D. degree at the 5% significance level.

- We perform the following steps:
  1. $\bar{W} = \frac{1}{n}\sum_{i=1}^{n} W_i = \frac{1}{250}\sum_{i=1}^{250} W_i = 61977.12$.
  2. $SE(\bar{W}) = \frac{s_W}{\sqrt{n}} = \frac{s_W}{\sqrt{250}} = 1334.19$.
  3. Compute $t^{act} = \frac{\bar{W} - \mu_{W,0}}{SE(\bar{W})} = \frac{61977.12 - 60000}{1334.19} = 1.4819$.
  4. Since we use a 5% significance level, we do not reject H_0 because $|t^{act}| = 1.4819 < z_{0.025} = 1.96$.
- Suppose we are interested in the alternative $H_1 : \mathrm{E}(W) > 60000$. The t-stat is **exactly** the same: $t^{act} = 1.4819$. but now needs to be compared with $z_{0.05} = 1.645$.

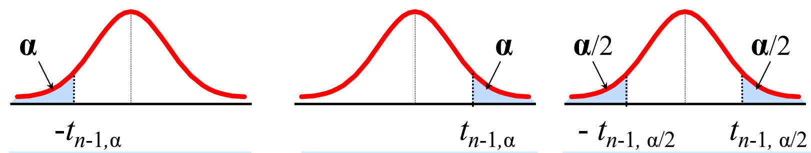
Hypothesis Test for the Mean with σ^2 **unknown** (n small)

Decision Rules

- Consider a random sample of n observations from a population that is normally distributed, ***AND*** variance σ_Y^2 is unknown: $Y_i \sim N(\mu_Y, \sigma_Y^2)$
- Converting the sample average ($\bar{y}$) to a t-value...

Hypothesis Tests for $E(Y)$ $\quad t = \frac{\bar{Y} - \mu_{Y,0}}{\mathrm{SE}(\bar{Y})} = \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} \sim t_{n-1}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(Y) \geq \mu_0$ H_1: $E(Y) < \mu_0$	H_0: $E(Y) \leq \mu_0$ H_1: $E(Y) > \mu_0$	H_0: $E(Y) = \mu_0$ H_1: $E(Y) \neq \mu_0$



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

Hypothesis Test for the Mean with σ^2 **unknown** (n small)

Example

- The average cost of a hotel room in New York is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{y} = \$172.50$ and $s_y = \$15.40$. Perform a test at the $\alpha = 0.05$ level (assuming the population distribution is normal).
  - Form the hypothesis to be tested

$$H_0 : \mathrm{E}(Y) = 168 \quad \text{the } \textit{mean} \text{ cost } \textbf{is \$168}$$
$$H_1 : \mathrm{E}(Y) \neq 168 \quad \text{the } \textit{mean} \text{ cost } \textbf{is not \$168}$$

  - For $\alpha = 0.05$, with $n = 25$, $t_{n-1,\alpha/2} = t_{24,0.025} = 2.0639$ and $-t_{24,0.025} = 2.0639$, so we would reject H_0 if $|t| > 2.0639$.
  - We have $\bar{y} = 172.50$ and $s_y = 15.40$, so the test statistic is

$$t = \frac{\bar{y} - \mu_{Y,0}}{s_y/\sqrt{n}} = \frac{172.50 - 168}{15.40/\sqrt{25}} = 1.46 < t_{24,0.025} = 2.0639$$

or $|t| = 1.46 < 2.0639$. Hence H_0 **cannot be** rejected. We **conclude** that there is not sufficient evidence that the true mean cost is different than \$168.

Confidence Intervals for the Population Mean – I

- Suppose we would do a two-sided hypothesis test for many different values of $\mu_{0,Y}$. On the basis of this we can construct a set of values which are not rejected at 5% (α%) significance level.
- If we were able to test all possible values of $\mu_{0,Y}$ we could construct a 95% ($(1-\alpha)$%) confidence interval

Definition

A 95% ($(1-\alpha)$%) confidence interval is an interval that contains the true value of μ_Y in 95% ($(1-\alpha)$%) of all possible random samples.

  - A relative frequency interpretation: From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true population mean

Confidence Intervals for the Population Mean – II

- The general formula for all confidence intervals is

$$\text{Point Estimate} \pm \underbrace{(\text{Reliability Factor})(\text{Standard Error})}_{\text{Margin of Error}}$$

$$\hat{\mu} \pm c \cdot \mathrm{SE}(\hat{\mu})$$

and using the sample average estimator

$$\bar{Y} \pm c \cdot \mathrm{SE}(\bar{Y})$$

- Instead of doing infinitely many hypothesis tests we can compute the 95% ($(1-\alpha)$%) confidence interval as

$$\bar{Y} - z_{\alpha/2}\mathrm{SE}(\bar{Y}) < \mu < \bar{Y} + z_{\alpha/2}\mathrm{SE}(\bar{Y}) \quad \text{or} \quad \bar{Y} \pm \underbrace{z_{\alpha/2}\mathrm{SE}(\bar{Y})}_{\text{Margin of Error}}$$

Confidence Intervals for the Population Mean – III

- When the sample size n is large (or when the population is normal and σ_Y^2 is known):
  - A 90% confidence interval for μ_Y: $[\bar{Y} \pm 1.645 \cdot \mathrm{SE}(\bar{Y})]$
  - A 95% confidence interval for μ_Y: $[\bar{Y} \pm 1.96 \cdot \mathrm{SE}(\bar{Y})]$
  - A 99% confidence interval for μ_Y: $[\bar{Y} \pm 2.58 \cdot \mathrm{SE}(\bar{Y})]$
  - with $\mathrm{SE}(\bar{Y}) = \sigma_Y/\sqrt{n}$ when variance is known or $\mathrm{SE}(\bar{Y}) = s_Y/\sqrt{n}$ when unknown and is estimated.

Confidence Intervals for the Population Mean – IV

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 0.35 ohms. Determine a 95% C.I. for the true mean resistance of the population.

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma_Y}{\sqrt{n}} = 2.20 \pm 1.96(0.35/\sqrt{11}) = 2.20 \pm 0.2068$$
$$1.9932 < \mu_Y < 2.4068$$

- 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

Confidence Intervals for the Population Mean – V

Example

Using the sample of $n = 250$ individuals with a Ph.D. degree discussed above ($\bar{W} = 61977.12, s_W = 21095.37, \text{SE}(\bar{Y}) = s_W/\sqrt{n} = 21095.37/\sqrt{250}$):

- A 90% C.I. for μ_W is: $[61977.12 \pm 1.64 \cdot 1334.19] = [59349.39, 64604.85]$.
- A 95% C.I. for μ_W is: $[61977.12 \pm 1.96 \cdot 1334.19] = [59774.38, 64179.86]$.
- A 99% C.I. for μ_W is: $[61977.12 \pm 2.58 \cdot 1334.19] = [58513.94, 65440.30]$.

Confidence Intervals for the Population Mean – VI

- When the sample size n is small ***AND*** the population from which we draw data is normal:

$$\bar{Y} - t_{n-1,\alpha/2}\frac{s_Y}{\sqrt{n}} < \mu_Y < \bar{Y} + t_{n-1,\alpha/2}\frac{s_Y}{\sqrt{n}} \quad \text{or} \quad \bar{Y} \pm \underbrace{t_{n-1,\alpha/2}\frac{s_Y}{\sqrt{n}}}_{\text{Margin of Error}}$$

- A 90% confidence interval for μ_Y: $[\bar{Y} \pm t_{n-1,0.05} \cdot \text{SE}(\bar{Y})]$
- A 95% confidence interval for μ_Y: $[\bar{Y} \pm t_{n-1,0.025} \cdot \text{SE}(\bar{Y})]$
- A 99% confidence interval for μ_Y: $[\bar{Y} \pm t_{n-1,0.005} \cdot \text{SE}(\bar{Y})]$
- with $\text{SE}(\bar{Y}) = s_Y/\sqrt{n}$

Confidence Intervals for the Population Mean – VII

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 $t_{24,\alpha/2} = t_{24,0.025} = 2.0639$

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = 50 \pm 2.0639(8/\sqrt{25}) = 50 \pm 3.302$$
$$46.698 < \mu < 53.302$$

Comparing Means from Different Populations – I

Large Samples or Known Variances from Normal Populations

- Suppose we would like to test whether the mean wages of men and women with a Ph.D. degree differ by an amount d_0:

$$H_0 : \mu_{W,M} - \mu_{W,F} = d_0 \quad H_0 : \mu_{W,M} - \mu_{W,F} \neq d_0$$

- To test the null hypothesis against the two-sided alternative we follow the 4 steps as above with some adjustments

1. Estimate $(\mu_{W,M} - \mu_{W,F})$ by $(\bar{W}_M - \bar{W}_M)$.
   - Because a weighted average of 2 independent normal random variables is itself normally distributed we have (using the CLT and the fact that $\text{Cov}(\bar{W}_M, \bar{W}_F) = 0$)

$$\bar{W}_M - \bar{W}_F \sim N\left(\mu_{W,M} - \mu_{W,F}, \frac{\sigma^2_{W,M}}{n_M} + \frac{\sigma^2_{W,F}}{n_F}\right)$$

Comparing Means from Different Populations – II

Large Samples or Known Variances from Normal Populations

2. Estimate $\sigma_{W,M}$ and $\sigma_{W,F}$ to obtain $\text{SE}(\bar{W}_M - \bar{W}_F)$:

$$\text{SE}(\bar{W}_M - \bar{W}_F) = \sqrt{\frac{s^2_{W,M}}{n_M} + \frac{s^2_{W,F}}{n_F}}$$

3. Compute the t-statistic

$$t^{act} = \frac{(\bar{W}_M - \bar{W}_M) - d_0}{\text{SE}(\bar{W}_M - \bar{W}_F)}$$

4. Reject H_0 at a 5% significance level if $|t^{act}| > 1.96$ or if the p-value< 0.05.

Comparing Means from Different Populations – III

Large Samples or Known Variances from Normal Populations

Example

Suppose we have random samples of 500 men and 500 women with a Ph.D. degree and we would like to test that the mean wages are equal:

$$H_0 : \mu_{W,M} - \mu_{W,M} = 0 \quad H_1 : \mu_{W,M} - \mu_{W,M} \neq 0$$

We obtained $\bar{W}_M = 64159.45$, $\bar{W}_F = 53163.41$, $s_{W,M} = 18957.26$, and $s_{W,F} = 20255.89$. We have:

1. $\bar{W}_M - \bar{W}_F = 64159.45 - 53163.41 = 10996.04$.
2. $\text{SE}(\bar{W}_M - \bar{W}_F) = 1240.709$.
3. $t^{act} = \frac{(\bar{W}_M - \bar{W}_F) - 0}{\text{SE}(\bar{W}_M - \bar{W}_F)} = \frac{10996.04}{1240.709} = 8.86$.
4. Since we use a 5% significance level, we reject H_0 because $|t^{act}| = 8.86 > 1.96$

Confidence Interval for the Difference in Population Means

- The method for constructing a confidence interval for 1 population mean can be easily extended to the difference between 2 population means.
- A hypothesized value of the difference in means d_0 will be rejected if $|t| > 1.96$ and will be in the confidence set if $|t| \leq 1.96$.
- Thus the 95% confidence interval for $\mu_{W,M} - \mu_{W,F}$ are the values of d_0 within ± 1.96 standard errors of $(\bar{W}_M - \bar{W}_F)$.
- So a 95% confidence interval for $\mu_{W,M} - \mu_{W,F}$ is

$$(\bar{W}_M - \bar{W}_M) \pm 1.96 \cdot \text{SE}(\bar{W}_M - \bar{W}_M)$$
$$10996.04 \pm 1.96 \cdot 1240.709$$
$$[8561.34, 13430.73]$$

Testing Population Mean Differences

Normal Populations, **Unknown Variances** σ_X^2 and σ_Y^2 but Assumed **Equal**

$$t = \frac{(\bar{X}-\bar{Y})-d_0}{\text{SE}(\bar{X}-\bar{Y})} = \frac{(\bar{X}-\bar{Y})-d_0}{\sqrt{(s_p^2/n_X)+(s_p^2/n_Y)}} \sim t_{n_X+n_Y-2};$$

$$\text{where } s_p^2 = \frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}$$

- The C.I. is constructed as $(\bar{X}-\bar{Y}) \pm t_{n_X+n_Y-2,\alpha/2} \cdot \text{SE}(\bar{X}-\bar{Y})$.
- Recall $\mu_X = \text{E}(X)$, $\mu_Y = \text{E}(Y)$

$H_0 : \mu_X - \mu_Y \geq d_0$ $H_1 : \mu_X - \mu_Y < d_0$	$H_0 : \mu_X - \mu_Y \leq d_0$ $H_1 : \mu_X - \mu_Y > d_0$	$H_0 : \mu_X - \mu_Y = d_0$ $H_1 : \mu_X - \mu_Y \neq d_0$
Lower-tail	Upper-tail	Two-tailed
Reject H_0 if $t < t_\alpha$	Reject H_0 if $t > t_\alpha$	Reject H_0 if $\lvert t \rvert > t_{\alpha/2}$

Testing Population Mean Differences – I

Example: Normal Populations, **Unknown Variances** σ_X^2 and σ_Y^2 but Assumed **Equal**

- 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 ($\alpha = 0.05$)?
  - The hypothesis of interest is

$$\begin{array}{l} H_0 : \mu_{NYSE} - \mu_{NASDAQ} = 0 \\ H_1 : \mu_{NYSE} - \mu_{NASDAQ} \neq 0 \end{array} \quad \text{or} \quad \begin{array}{l} H_0 : \mu_{NYSE} = \mu_{NASDAQ} \\ H_1 : \mu_{NYSE} \neq \mu_{NASDAQ} \end{array}$$

Testing Population Mean Differences – II

Example: Normal Populations, **Unknown Variances** σ_X^2 and σ_Y^2 but Assumed **Equal**

- Note that $df = n_X + n_Y - 2 = 21 + 25 - 2 = 44$, so the critical value for the test is $t_{44,0.025} = 2.0154$
- The pooled variance is:

$$s_p^2 = \frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2} = \frac{(21-1)1.30^2+(25-1)1.16^2}{(21-1)+(25-1)} = 1.5021$$

- The test statistic is

$$t^{act} = \frac{(\bar{x}-\bar{y})-d_0}{\sqrt{(s_p^2/n_X)+(s_p^2/n_Y)}} = \frac{(3.27-2.53)-0}{\sqrt{1.5021\left(\frac{1}{21}+\frac{1}{25}\right)}} = 2.040.$$

  Since $\lvert t^{act} \rvert > t_{44,0.025} = 2.0154$, we reject H_0 at $\alpha = 0.05$. We conclude that there is evidence of a difference...

- The C.I. is constructed as $(\bar{X}-\bar{Y}) \pm t_{n_X+n_Y-2,\alpha/2} \cdot \text{SE}(\bar{X}-\bar{Y})$

Testing Population Mean Differences – I

Matched or Paired Samples

- Suppose we obtain a sample of n observations from two populations which are normally distributed and we have paired or matched samples – repeated measures (before/after).
- Define, the pair difference $d_i = X_i - Y_i$. We have

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{X} - \bar{Y}; \quad \text{and} \quad S_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2}$$

  with $\text{E}(\bar{d}) = \mu_d = \text{E}(X) - \text{E}(Y)$ and $\text{SE}(\bar{d}) = \sqrt{\frac{S_d^2}{n}} = S_d/\sqrt{n}$
- If the sample size is large enough ($n \to \infty$) then

$$\frac{\bar{d}-\mu_d}{S_d/\sqrt{n}} \sim N\left(0, \frac{S_d^2}{n}\right).$$

  If the sample size is relatively small, then

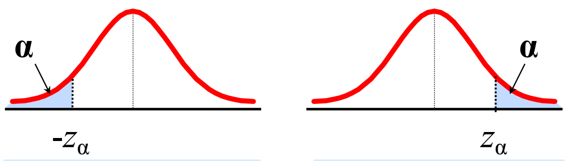
$$\frac{\bar{d}-\mu_d}{S_d/\sqrt{n}} \sim t_{n-1}.$$

Testing Population Mean Differences – II

Matched or Paired Samples

Matched or Paired Samples $t = \frac{\bar{d} - d_0}{\mathrm{SE}(d)} = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}$ (n large)

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(X)–E(Y) \geq 0$	H_0: $E(X)–E(Y) \leq 0$	H_0: $E(X)–E(Y) = 0$
H_1: $E(X)–E(Y) < 0$	H_1: $E(X)–E(Y) > 0$	H_1: $E(X)–E(Y) \neq 0$



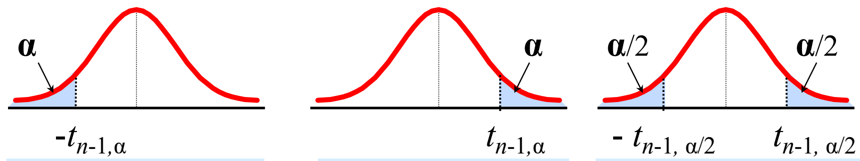
Reject H_0 if $t < -z_\alpha$	Reject H_0 if $t > z_\alpha$	Reject H_0 if $t < -z_{a/2}$ or $t > z_{a/2}$

Testing Population Mean Differences – III

Matched or Paired Samples

Matched or Paired Samples $t = \frac{\bar{d} - d_0}{\mathrm{SE}(d)} = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} \sim t_{n-1}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(X)–E(Y) \geq 0$	H_0: $E(X)–E(Y) \leq 0$	H_0: $E(X)–E(Y) = 0$
H_1: $E(X)–E(Y) < 0$	H_1: $E(X)–E(Y) > 0$	H_1: $E(X)–E(Y) \neq 0$



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

Testing Population Mean Differences – I

Matched or Paired Samples: Example

- Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? Test at the 5% significance level. You collect the following data:

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

$$\bar{d} = \frac{1}{5}\sum_{i=1}^{5} d_i = -4.2;\ s_d = \sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(d_i - \bar{d})^2} = 5.67$$

- The hypothesis of interest is

$$H_0 : \mu_X - \mu_Y = 0$$
$$H_1 : \mu_X - \mu_Y \neq 0$$

Testing Population Mean Differences – II

Matched or Paired Samples: Example

- With $n = 4$ and $\alpha = 0.05$ the critical value is $t_{n-1,\alpha/2} = t_{4,0.025} = 2.776$.
- We have

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{4}} = -1.66 > -t_{4,0.025} = -2.776,$$

or $|t| < t_{4,0.025} = 2.776$. Hence, we **do not reject** H_0. There is not a significant change in the number of complaints.

Annex: Hypothesis Tests – I

Employing the p-value

- Suppose we have a sample of n observations (they are assumed *IID*) and compute the sample average $\bar{Y}$. The sample average can differ from $\mu_{Y,0}$ for two reasons
  1. The population mean μ_Y is not equal to $\mu_{Y,0}$ (H_0 is not true)
  2. Due to random sampling $\bar{Y} \neq \mu_Y = \mu_{Y,0}$ (H_0 is true)
- To quantify the second reason we define the p-value. The ***p*-value** is the probability of drawing a sample with $\bar{Y}$ at least as far from $\mu_{Y,0}$ as the value actually observed, given that the null hypothesis is true.

$$p\text{-value} = \Pr_{H_0}\left[\left|\bar{Y} - \mu_{Y,0}\right| > \left|\bar{Y}^{act} - \mu_{Y,0}\right|\right],$$

where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed

Annex: Hypothesis Tests – II

Employing the p-value

- To compute the p-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if n is small. With large n the CLT states that

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right),$$

which implies that if the null hypothesis is true:

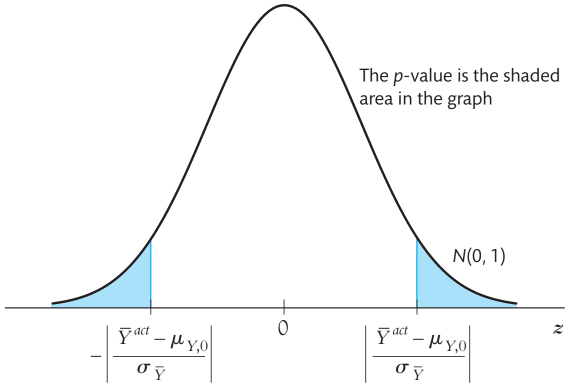
$$\frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \sim N(0,1)$$

- Hence

$$p\text{-value} = \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}}\right|\right] = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}}\right|\right)$$

Annex: Hypothesis Tests – III

Employing the p-value



- For large n, p-value = the probability that a $N(0,1)$ random variable falls outside $\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|$, where $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$

Annex: Hypothesis Tests – I

Computing the p-value when σ_Y^2 is unknown

- In practice σ_Y^2 is usually unknown and must be estimated
- The sample variance S_Y^2 is the estimator of $\sigma_Y^2 = \mathrm{E}\left[(Y - \mu_Y)^2\right]$, defined as

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  - division by $n-1$ because we 'replace' μ_Y by $\bar{Y}$ which uses up 1 degree of freedom
  - if $Y_1, Y_2, \ldots, Y_n$ are *IID* and $\mathrm{E}(Y^4) < \infty$, then $S_Y^2 \xrightarrow{p} \sigma_Y^2$ (Law of Large Numbers)
- The sample standard deviation $S_Y = \sqrt{S_Y^2}$, is the estimator of σ_Y.

Annex: Hypothesis Tests – II

Computing the p-value when σ_Y^2 is unknown

- The standard error $SE(\bar{Y})$ is an estimator of $\sigma_{\bar{Y}}$

$$SE(\bar{Y}) = \frac{S_Y}{\sqrt{n}}$$

- Because S_Y^2 is a consistent estimator of σ_Y^2 we can (for large n) replace

$$\sqrt{\frac{\sigma_Y^2}{n}} \text{ by } SE(\bar{Y}) = \frac{S_Y}{\sqrt{n}}$$

- This implies that when σ_Y^2 is unknown and $Y_1, Y_2, ..., Y_n$ are *IID* the p-value is computed as

$$p - \text{value} = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}\right|\right)$$