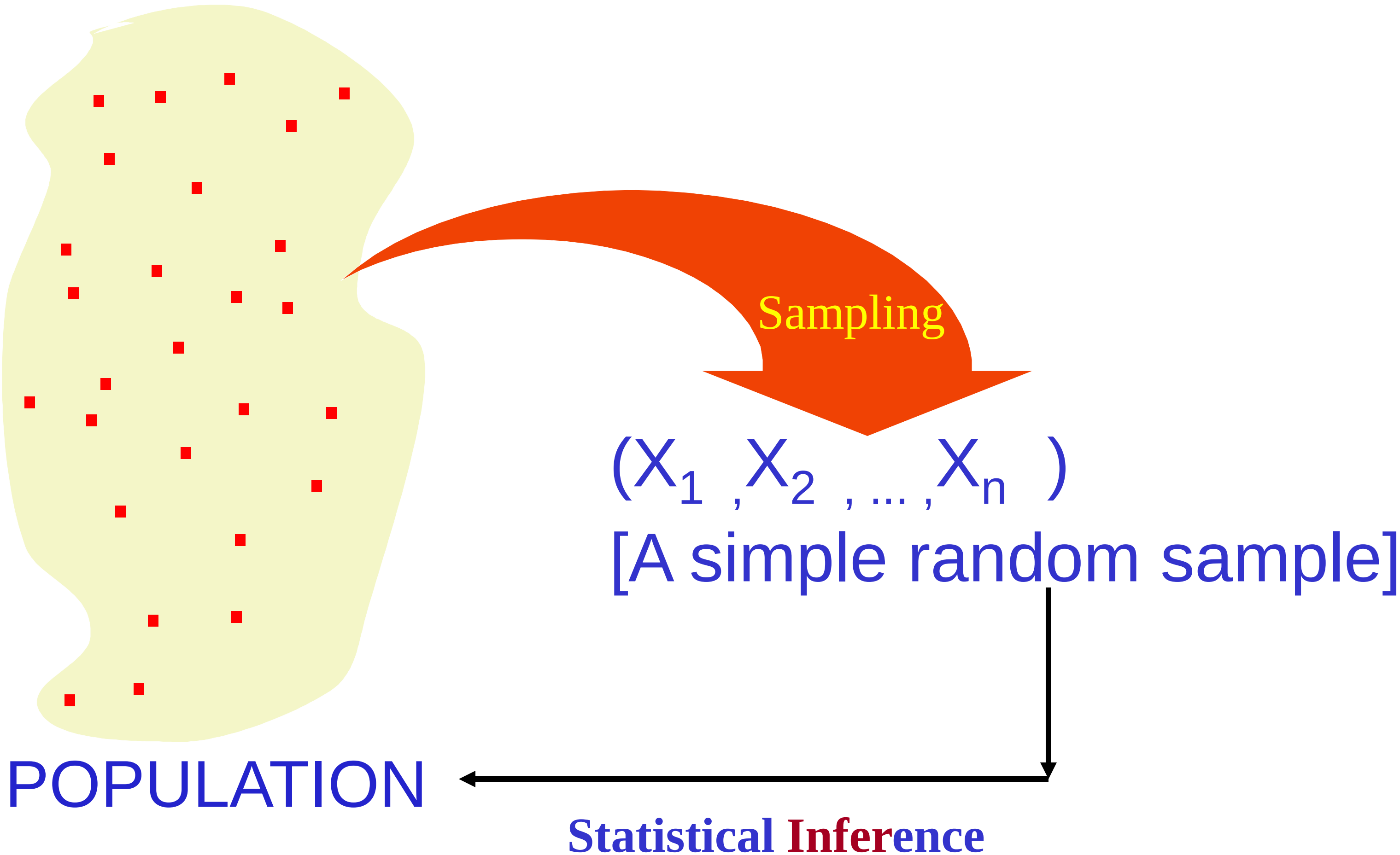
Review of Statistics

- The reality we don't know the population (e.g., and σ in the normal population).



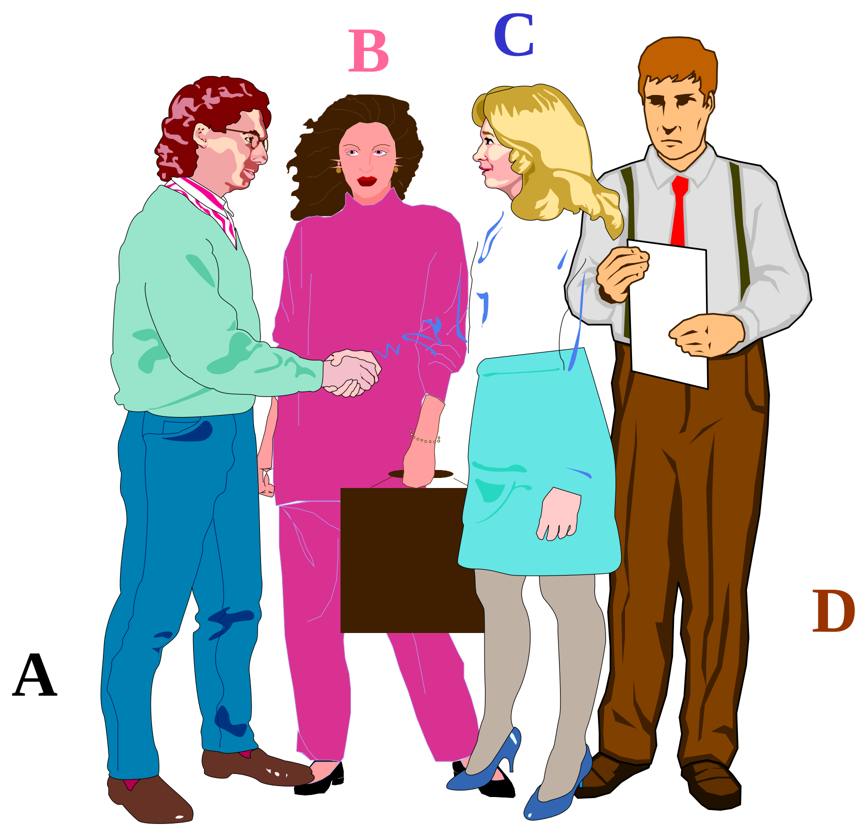
An Example

- Suppose the true population has four people.
- You are interested in their age
- A: age 18

  B: age 20

  C: age 22

  D: age 24
- Suppose that the true population is unknown, so we randomly survey people.
- (Of course, this is just a simple example. In practice, the population can be very large)



An Example

- We randomly survey one of the four people and record the person's age as X1. So X1 is the first data.
- X1 is random and X1 follow the distribution of the population:

$$X_1 = \begin{cases} \text{18 with probability 1/4} \\ \text{20 with probability 1/4} \\ \text{22 with probability 1/4} \\ \text{24 with probability 1/4} \end{cases}$$

- We survey the second people and record the person's age as X2. So X2 is the second data.
- X2 is random and X2 follow the same distribution of the population.
- X1 and X2 are independent.
- We collect n data, {X1,X2,…Xn}. These n data are n random variables.
- These n data follow the identical distribution of the population.
- These n data are independent.

- Probability: use information from populations to learn about samples

- Statistics: use information from samples to learn about populations (population parameters are unknown numbers)

- In statistics:
    - Estimation
    - Hypothesis testing
    - Confidence interval

An Example of coin-flipping

- I have an "unfair" coin:

$$\begin{cases} 1\ (head) \text{ with probability } p \\ 0\ (tail) \quad \text{with probability } 1-p \end{cases}$$

- p is an unknown number, which is not necessarily equal to ½ ; for example, p can be 0.1, 0.2….or any number between 0 and 1.
- I am interested in the unknown number p (population parameter).
- So I flip the unfair coin 5 times (my sample) and record the outcomes for each time as:

Y_1 : the outcome for the first flipping,

Y_2 : the outcome for the second flipping

…..

Y_5 : the outcome for the 5th flipping

An Example

- Some clarifications:
  - $\{Y_1, Y_2, \ldots, Y_5\}$ is 5 data points; these data are random variable!
  - Sample average $\bar{Y}$ is a random variable

    e.g. $\{Y_1, Y_2, \ldots, Y_5\}$ can be {1,0, 1,0,1} with $\bar{Y}$=3/5

    $\{Y_1, Y_2, \ldots, Y_5\}$ can be {0,0,1,1,0} with $\bar{Y}$=2/5.
  - $\{Y_1, Y_2, \ldots, Y_5\}$ are iid from the unknown population distribution:

$$\begin{cases} 1\ (head) \text{ with probability } p \\ 0\ (tail) \quad \text{with probability } 1-p \end{cases}$$

•Why iid?

•Independent: because the first coin-flipping has nothing to do with the second coin-flipping…

•Identically distributed: because each of $\{Y_1, Y_2, \ldots, Y_5\}$ follows the same unknown distribution as the population:

$$\begin{cases} 1\ (head) \text{ with probability } p \\ 0\ (tail) \quad \text{with probability } 1-p \end{cases}$$

An Example

- Population

$$\begin{cases} 1\ (head) \text{ with probability } p \\ 0\ (tail) \text{ with probability } 1-p \end{cases}$$

- Sample: Random variables

$$\{Y_1, Y_2, \ldots, Y_5\}$$

Before we see the realization of the data

- Realizations

{Head, Tail, Head, Head, Tail}

An Example

- $\{Y_1, Y_2, \ldots, Y_5\}$ iid
- What is an estimator of p? Naturally, you think about the sample average! But how good is this estimator?
- Hypothesis testing: for example, I am interested if the coin is fair, i.e., is the true p=1/2. Can we test it using data? Well, we can calculate the sample average and to see if it is equal to ½ .
  - If it turns out that the data $\{Y_1, Y_2, \ldots, Y_5\}$= {1,0, 1,0,1}, then it gives the sample average=3/5. Can we reject p=1/2?
  - No!!! Because even if the true p=1/2, it is still possible to observe {1,0, 1,0,1}.
  - $\bar{Y}$ is random!!! Even it turns out that the sample average =3/5. This could be completely due to randomness of $\bar{Y}$.
- Confidence interval: using an interval to estimate p.

1) Estimation of Population Mean

- From a random sample, we have the data $\{Y_1, Y_{2,\ldots,} Y_n\}$, (for example income level)
- It is from underlying distribution with population mean mu (μ).
- How can we estimate mu? Average, 1st observation, median?
- It turns out that the sample average $\bar{Y}$ is desirable
- What makes a statistical estimator "desirable"
  - Properties (i) unbiased (expected value)
    (ii) consistent (as n goes infinite)
    (iii) efficient (small variance)

1) Estimation of Population Mean

- Bias: An estimator $\hat{\theta}$ has bias defined as

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta,$$ where θ is population parameter.

- If Bias=0, then $\hat{\theta}$ is unbiased estimator of θ .

i.e., what the average of $\hat{\theta}$ from many repeated sample? hopefully, it's θ .

- $\bar{Y}$ is unbiased estimator of population mean μ .

$$\bar{Y} = \frac{1}{n}\sum Y_i$$

$$E(\bar{Y}) = E\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n}\cdot n\mu = \mu$$

$$Bias(\bar{Y}) = E(\bar{Y}) - \mu = \mu - \mu = 0,$$ unbiased!

1) Estimation of Population Mean

- Consistency: $\hat{\theta}$ is a consistent estimator for population parameter θ if

$$\hat{\theta} \xrightarrow{p} \theta,$$

$$i.e., \ \Pr\left(\left|\hat{\theta} - \theta\right| < c\right) \rightarrow 1, \text{ as } n \rightarrow \infty$$

$$or, \quad \Pr\left(\left|\hat{\theta} - \theta\right| > c\right) \rightarrow 0, \text{ as } n \rightarrow \infty$$

- $\bar{Y}$ is a consistent estimator of μ. (Proof by Law of Large Number !)
- Law of Large Number says that $\bar{Y} \xrightarrow{p} \mu$.

1) Estimation of Population Mean

- Efficiency: Let $\tilde{\theta}$ be another unbiased estimator of population parameter θ .

$$\hat{\theta} \text{ is more efficient than } \tilde{\theta} \text{ if } var\left(\hat{\theta}\right) < var(\tilde{\theta})$$

- We can show that $var(\bar{Y}) = \frac{\sigma^2}{n}$ < variance of any other unbiased linear estimator.

- So $\bar{Y}$ is "BLUE" (Best Linear Unbiased Estimator)

- Example: another unbiased estimator: $\tilde{Y} = \frac{1}{2}Y_1 + \frac{1}{2}Y_2$

  $\bar{Y}$ is more efficient than $\tilde{Y}$, as

$$var(\tilde{Y}) = \frac{1}{4}var(Y_1) + \frac{1}{4}var(Y_2) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{1}{2}\sigma^2 > \frac{1}{n}\sigma^2 = var(\bar{Y})$$

1) Estimation of Population Mean

- Population variance σ^2 is also unknown usually.
- An estimator for σ^2 is sample variance

$$s^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

- This is unbiased and consistent estimator for σ^2.
- Note that the sample standard deviation:

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

  is estimator for standard deviation σ.
- The standard error of $\bar{Y}$ is an estimator of standard deviation of $\bar{Y}$. Remember the standard deviation of $\bar{Y}$ is $\sqrt{\frac{\sigma^2}{n}}$. Thus its estimator, standard error is $\sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$.

2) Hypothesis Testing

- We don't know the true population mean parameter
- We want to test if

$$H_0 : \mu = \mu^*$$

where μ^* is some constant (for example, μ^* =3).

- E.g. we have a sample of 1000 individuals with mean income and sample standard deviation:

$$\bar{Y} = 57557.7; \quad s = 59806.6$$

Test the hypothesis the true population mean income (μ) is 60000 (μ^*) .

- Set up the hypothesis:

$$(Null)\ H_0 : \ \mu = \mu^*$$

$$(Alternative)\ H_1 : \mu \neq \mu^*$$

2) Hypothesis Testing

- We can use at least 3 methods to reject or not reject H_0.

  (i) t-statistics (ii) p-value

  (iii) confidence interval
- In any case, first need to choose a <u>significance level</u> (α), usually either 1%, 5% or 10%.
- Type I error: Reject Null when it is true

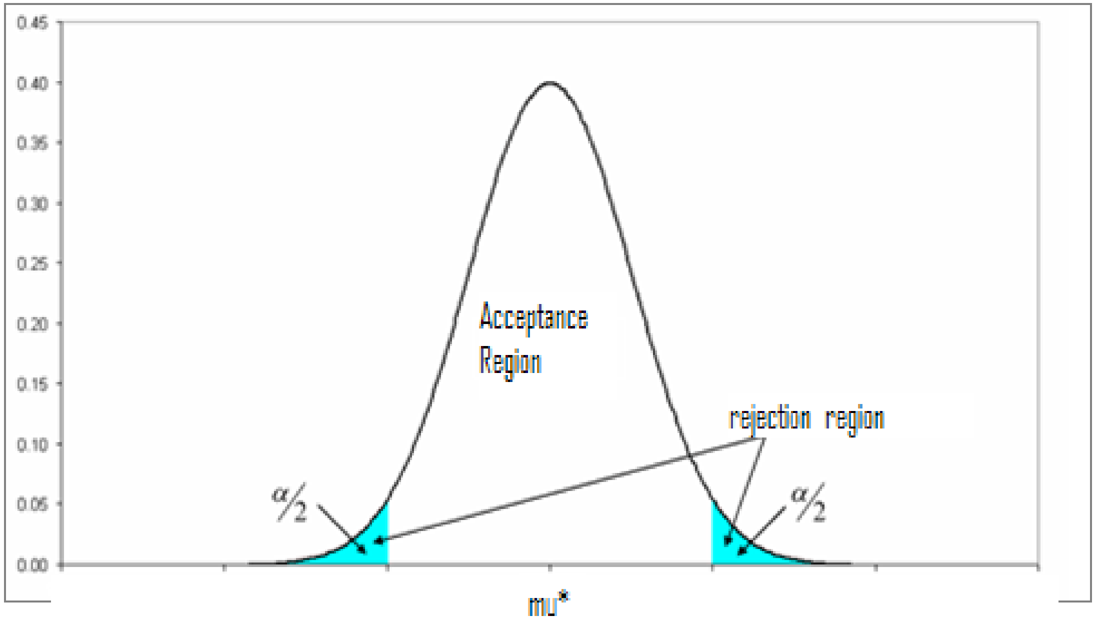
  Type II error: Not reject Null when it is false
- Significance level: a pre-specified rejection probability of a hypothesis test when null is true. (i.e., it is prob(Type I error))
- Type I and Type II error are unavoidable!
- Typically, we pre-specify Type I error and try to minimize the Type II error.

2) Hypothesis Testing

- Intuitively, if $\bar{Y}$ is sufficiently far away from μ^* , then we should regret H_0.
- In coin-flipping example,
  - suppose that you flip coin 1000 times and the sample average =0.98; this provide "significant" evidence that we should reject the hypothesis that the coin is fair (p=1/2).
  - Suppose that you flip coin 1000 times and the sample average =0.52; this does Not provide "significant" evidence that we should reject the hypothesis that it is a fair coin (p=1/2).
- How far is far away enough? We need to quantify the uncertainty of the statistic $\bar{Y}$.

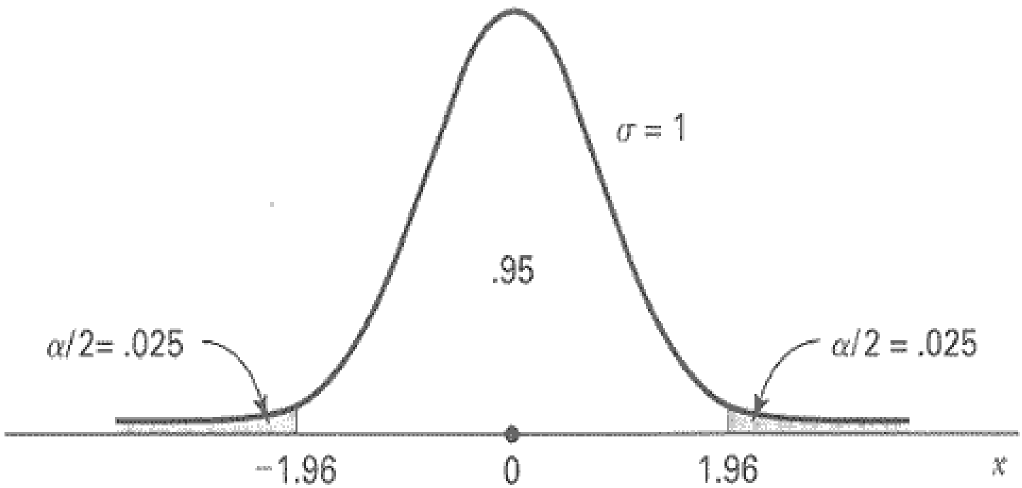
2) Hypothesis Testing

- If sample size is large, from central limit theorem, one we can utilize normal distribution.
- Let's pick significance level of 5% (0.05)
- Two sided hypothesis test: alternative H_1, is $\mu \neq \mu^*$
- Suppose μ^*=100, $\alpha = 0.05$. Then under the null, $\bar{Y}$ should be central around μ^*:



2) Hypothesis Testing

- Normalized t-stat (suppose the population standard deviation σ is unknown): $t = \dfrac{\bar{Y} - \mu^*}{s/\sqrt{n}}$
- Under the null: using CLT we can show that t follows standard normal distribution



- If $t = \dfrac{\bar{Y} - \mu^*}{s/\sqrt{n}} \geq 1.96$ or ≤ -1.96, then $\bar{Y}$ is sufficiently far away form μ^* in order to reject $H_0 : \mu = \mu^*$

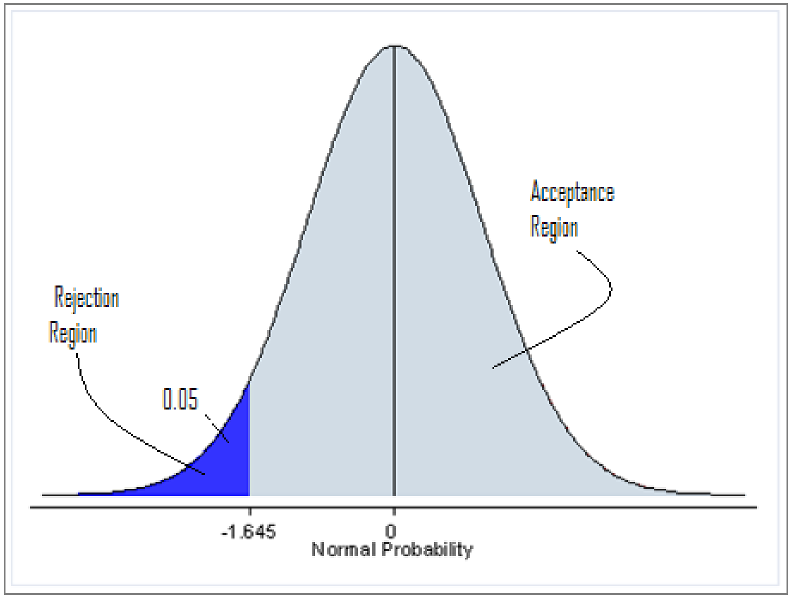
2) Hypothesis Testing

- One-sided test:

$$H_0 : \mu = \mu^* \text{ v.s. } H_1 : \mu < \mu^* \text{ (one sided, <)}$$

$$H_0 : \mu = \mu^* \text{ v.s. } H_1 : \mu > \mu^* \text{ (one sided, >)}$$

- Suppose significance level α =0.05.
  - For one-sided (<) test,
    if $t = \frac{\bar{Y} - \mu^*}{s/\sqrt{n}} < -1.65$, we reject
  - For one-sided (>) test,
    if $t = \frac{\bar{Y} - \mu^*}{s/\sqrt{n}} > 1.65$, we reject



2) Hypothesis Testing

- Example: US income

  $n = 1000$ people, $\bar{y} = \$57557.7$, $s = 59806.6$
- We are pre-specifying significance level $\alpha = 0.05$ (our tolerance for type I error is 0.05)
- We are testing

$$\begin{aligned} H_0 &: \mu = 60000 \\ H_1 &: \mu \neq 60000 \end{aligned}$$

- t-stat:

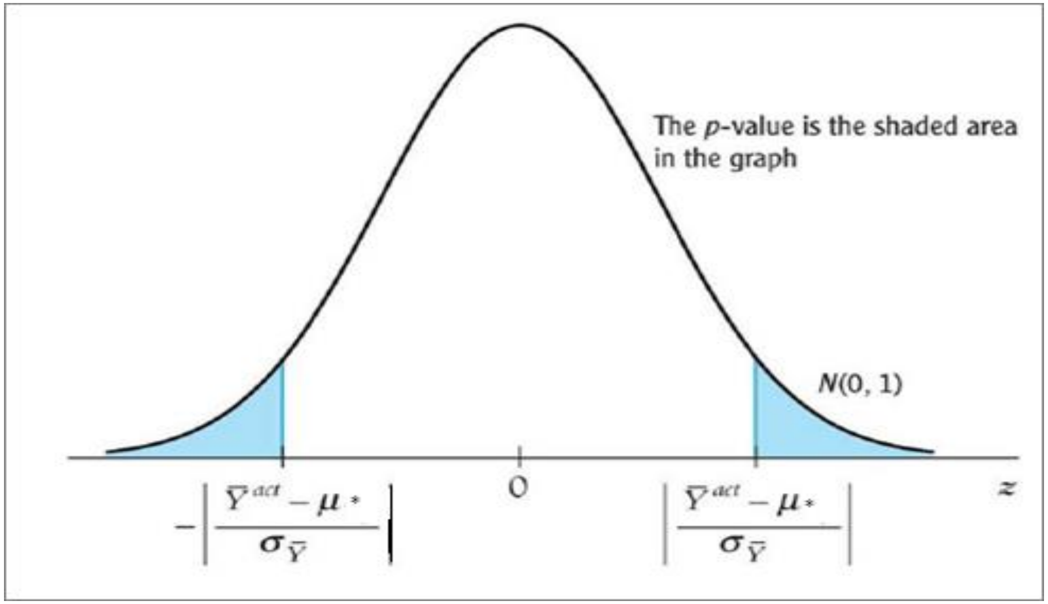
$$t_{act} = \frac{\bar{Y}_{act} - \mu^*}{s/\sqrt{n}} = \frac{57557.7 - 60000}{59806.6/\sqrt{1000}} = -1.29$$

- Clearly, -1.29 is between -1.96 and 1.96, thus we cannot reject H_0.

2) Hypothesis Testing

- p-value way:

  the p-value is probability of obtaining an $\bar{Y}$ value different from μ due to sampling variation given that we have observed $\bar{Y}_{act}$.
- If p-value is "large", we don't reject H_0 .
- Suppose we are testing: $H_0 : \mu = \mu^*$, v.s. $H_1 : \mu \neq u^*$

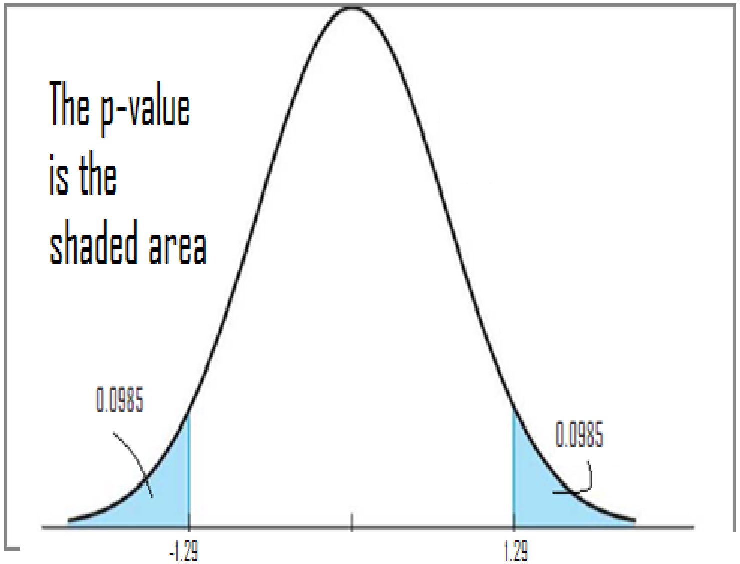


2) Hypothesis Testing

- 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.
- Use α value for significance. If p-value $< \alpha$, then reject H_0.
- Our example:

$$\begin{aligned} H_0 &: \mu = 60000 \\ H_1 &: \mu \neq 60000 \end{aligned}$$

- Use normal distribution, t=-1.29.
  p-value=2*0.0985=0.197
- p-value>0.05 so we cannot reject the null at α.



3) Confidence Interval (CI)

- Here, we are going to come up with a range of numbers (an interval) which will contain the population parameter μ.

- $(1-\alpha)\times 100\%$ of the time that the true μ will be on the interval in repeated samples; i.e, we want to find two random variables k_1 and k_2 such that

$$\Pr(k_1 < \mu < k_2) = 1 - \alpha$$

3) Confidence Interval (CI)

- We want to find two random variables k_1 and k_2 such that

$$\Pr(k_1 < \mu < k_2) = 1 - \alpha$$

- For a large sample, $\frac{\bar{Y}-\mu}{s/\sqrt{n}} \overset{A}{\sim} N(0,1)$,

$$\text{thus } \Pr\left(-z_{\frac{\alpha}{2}} < \frac{\bar{Y}-\mu}{s/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Rightarrow \Pr\left(-z_{\frac{\alpha}{2}} \bullet \frac{s}{\sqrt{n}} < \bar{Y} - \mu < z_{\frac{\alpha}{2}} \bullet \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Rightarrow \Pr\left(\bar{Y} - z_{\frac{\alpha}{2}} \bullet \frac{s}{\sqrt{n}} < \mu < \bar{Y} + z_{\frac{\alpha}{2}} \bullet \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

- If α=0.05, $z_{\frac{\alpha}{2}} = 1.96$. Thus the 95% CI for μ is

$$\left[\bar{Y} - 1.96\frac{s}{\sqrt{n}}, \bar{Y} + 1.96\frac{s}{\sqrt{n}}\right]$$

3) Confidence Interval

- Construct the confidence interval for our example:

$$\bar{Y} = 57557.7;\ s = 59806.6;\ n = 1000$$

- So the confidence interval is

$$\left[57557.7 - 1.96 \cdot \frac{59806.6}{\sqrt{1000}},\ 57557.7 + 1.96 \cdot \frac{59806.6}{\sqrt{1000}}\right]$$

$$= \left[53846,\ 61269\right]$$

3) Confidence Interval

- We can use confidence interval to test hypothesis:

$$H_0 : \mu = 60000$$
$$H_1 : \mu \neq 60000$$

- If μ^* falls inside the confidence interval, then we cannot reject the null $H_{0.}$

- If μ^* does not fall inside the confidence interval, then we reject the null

- In our example, μ^*=6000 is within the confidence interval: $\left[53846,\ 61269\right]$. Again, we fail to reject the Null.

4) Difference in Means

- Suppose that we want to test whether men's salary is different from women's salary by a certain amount
- Information we have:

$$Men : \bar{Y}_M = 63824.6,\ s_M = 58479.0,\ n_M = 870$$

$$Women : \bar{Y}_W = 56960.5,\ s_W = 59928.7,\ n_W = 913$$

- We want to test

$$H_0 : \mu_M - \mu_W = d^*$$

$$H_1 : \mu_M - \mu_W \neq d^*$$

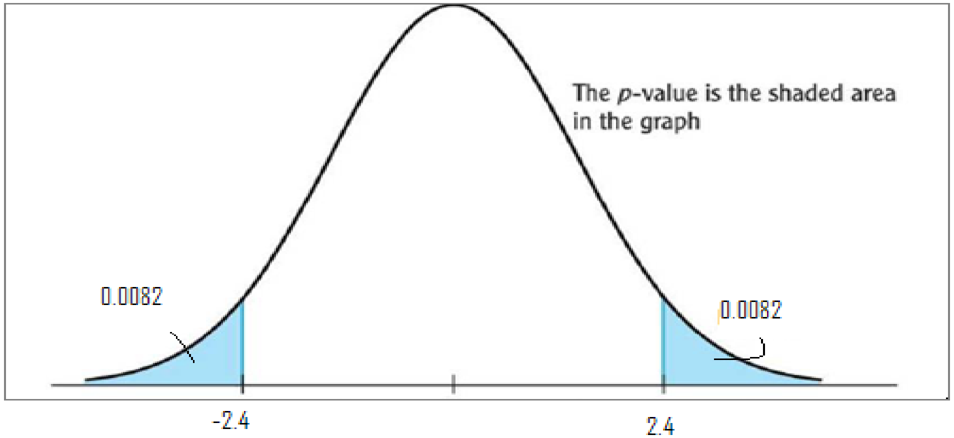
- Same procedure as before except now

$$t = \frac{(\bar{Y}_M - \bar{Y}_W) - d^*}{SE(\bar{Y}_M - \bar{Y}_W)}, \text{ where } SE(\bar{Y}_M - \bar{Y}_W) = \sqrt{\frac{s_M^2}{n_M} + \frac{s_W^2}{n_W}}$$

- Here $t = \frac{63824.5 - 56960.5 - 0}{\sqrt{\frac{58479.0^2}{870} + \frac{59928.7^2}{913}}} = 2.4 > 1.96$, so we reject.

4) Difference in Means

- Or we can use p-value=0.0082+0.0082=0.0164<0.05, we reject .



- CI:

$$\left[(\bar{Y}_M - \bar{Y}_W) - z_{\frac{\alpha}{2}} \cdot SE,\ \ (\bar{Y}_M - \bar{Y}_W) + z_{\frac{\alpha}{2}} \cdot SE\right]$$

$$= \left[63824.5 - 56960.5 - 1.96 \cdot \sqrt{\frac{58479.0^2}{870} + \frac{59928.7^2}{913}},\ \ 63824.5 - 56960.5 + 1.96 \cdot \sqrt{\frac{58479.0^2}{870} + \frac{59928.7^2}{913}}\right]$$

$$= \left[1400,\ 12000\right]$$

- 0 is not in the interval, so we reject $H_0 : \mu_M - \mu_W = 0$.

5) Treatment effects

- We can apply the difference of means to estimate causal/treatment effect using experiment data

- Take the difference of mean of treatment group and mean of control group

- The statistical procedure is exactly the same as in section (4)

6) Small Sample

- If sample size is small and population variance is unknown (as is usually the case), then should use t distribution instead of the normal (Normal is fine in large samples due to central limit theorem).

- So critical value should come from student t distribution with n-1 degrees of freedom.

6) Small Sample

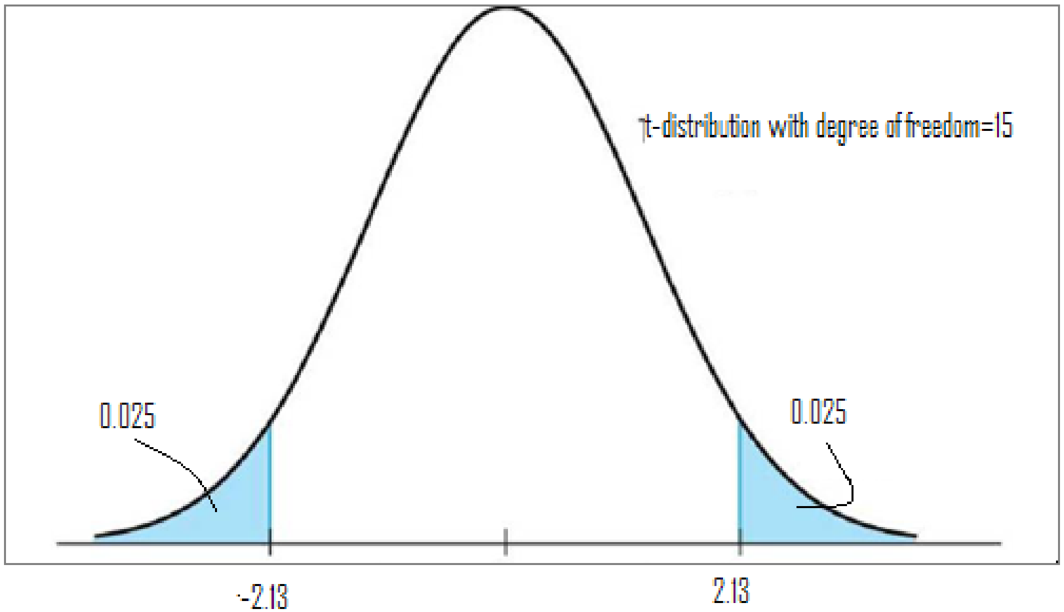
- Example: Income $\bar{Y} = 40,000, \quad s = 50,000, \quad n = 16$

  Want to test: $H_o : \mu = 44,000$ v.s. $H_1 : \mu \neq 44,000$

  Suppose the significance level is 0.05.

- $t = \frac{40,000-44,000}{\frac{50,000}{\sqrt{16}}} = -3.20$, reject since -3.20<2.13,

- 2.13 is from t-distribution with degree of freedom 15.



7) Covariance and Correlation

- Given a random sample of 2 variables, X and Y, we can estimate the relationship between them.
- For i=1, 2, …, n, we get (X_i, Y_i),
- How do X and Y covary?
- Sample covariance (an estimator of population covariance):

$$S_{XY} = \frac{1}{n-1}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

- If $S_{XY}>0$, then X and Y move together (e.g., income and education)
- If $S_{XY} <0$, then X and Y move opposite
- If $S_{XY}=0$, then no (linear) relationship

7) Covariance and Correlation

- Sample correlation (an estimator of population correlation):

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{1}{n-1}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum (X_i - \bar{X})^2}\sqrt{\frac{1}{n-1}\sum (Y_i - \bar{Y})^2}}$$

- We can show $-1 \le r_{XY} \le 1$

- Correlation gives direction and strength of linear relationship