# Outline

# Ejemplo: Proceso de Control Estadístico I

- Sea $X$ la duración de una llamada atendida por el agente del call center. Supongamos que$X$ es una v.a. que se distribuye $X \sim N(\mu_0, \sigma^2)$.
- Queremos testear si la duración de una llamada es mayor en la semana de vuelta de las vacaciones. Nuestra intuición es que las llamadas son, en promedio, un 20% mas largas. *Para ello formulamos dos hipótesis:*
  - Hipótesis Nula: $H_0 : \mu = \mu_0$(por ejemplo, 4 mins)
  - Hipótesis Alternativa: $H_1 : \mu = \mu_1(\text{por ejemplo, } 4.4 \text{ mins})$
- Asumamos además que $\sigma$ es conocida y es la misma independiente de la hipótesis.
- Nuestro objetivo es testear si es $H_0$ o $H_1$ la que mejor se ajusta a nuestros datos. Hacemos esto en tres pasos:
  1. Obtenemos la data: una muestra de $n$ observaciones, $x_1, \cdots, x_n$

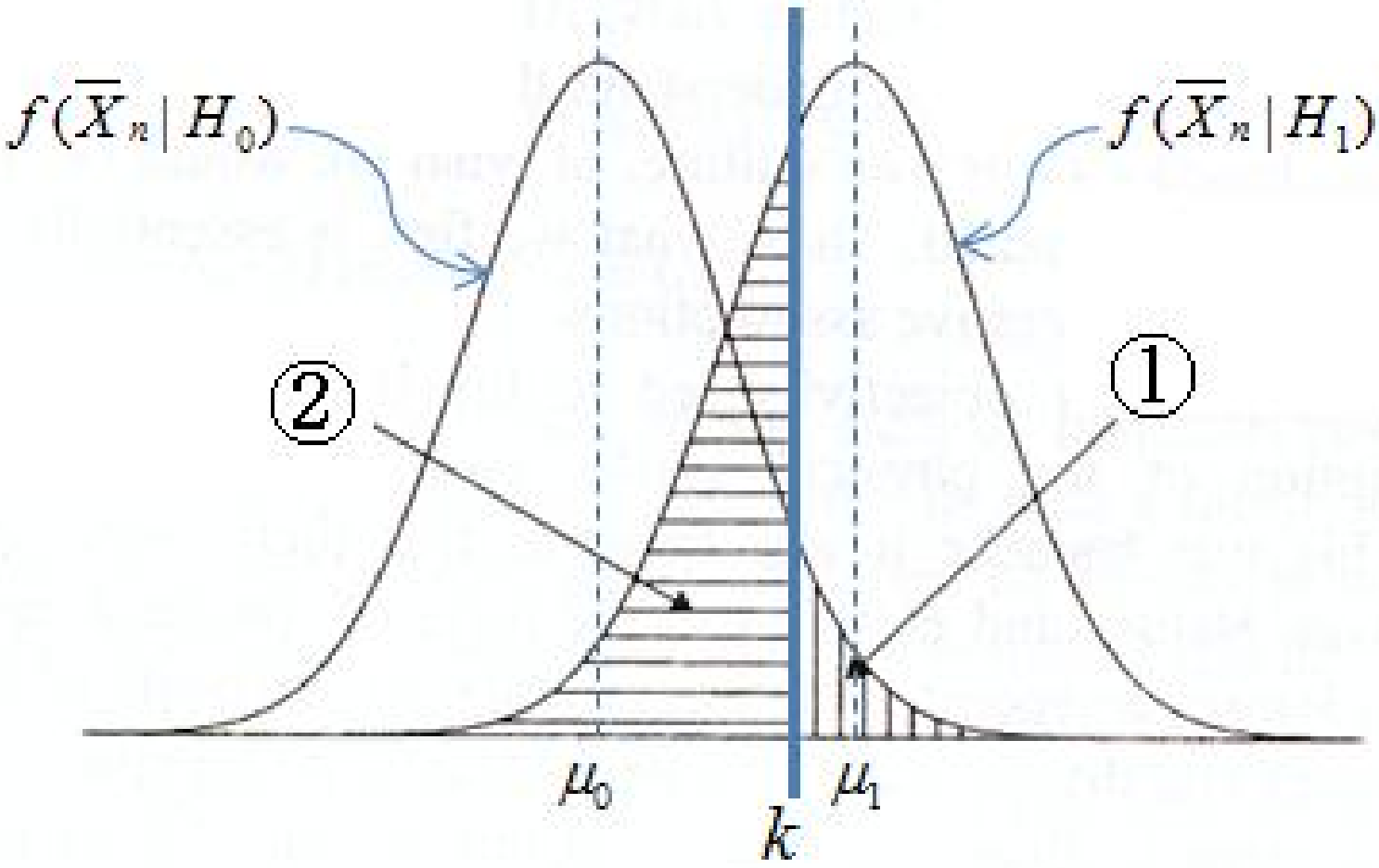
# Ejemplo: Proceso de Control Estadístico II

2. Calculamos el estadístico de interés: en nuestro ejemplo, la media muestral, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
3. Obtenemos una conclusión: Si $\overline{X}_n$ es demasiado grande, por ejemplo, $\overline{X}_n > k$ (una constante), *entonces rechazamos* $H_0$ "en favor de $H_1$" (i.e. preferimos $H_1$ sobre $H_0$). De lo contrario, *"no tenemos suficiente evidencia para afirmar que* $H_0$ es falsa".

- Cuando rechazamos $H_1$, estamos aceptando $H_0$? NO. El test de significancia sólo nos permite desacreditar una cierta hipótesis, pero no necesariamente confirmar la alternativa

# Ejemplo: Proceso de Control Estadístico III

Figure: Dos tipos de error



# Error Tipo I y Error Tipo II

Cuando hacemos test de hipótesis, hay dos tipos de errores que podemos cometer:

- Error Tipo I: Observamos que $\overline{X}_n > k$ , sin embargo la distribución verdadera del estadístico es $f(\overline{X}_n|H_0)$. En consecuencia,

$$\begin{aligned} \text{Error Tipo I} &= \Pr(\overline{X}_n > k|H_0) = \Pr(\textit{Rechazar } H_0|H_0 \textit{ verdadera}) \\ &= \text{Area 1} \end{aligned}$$

- Error Tipo II: Observamos que $\overline{X}_n < k$ , sin embargo la distribución verdadera del estadístico es $f(\overline{X}_n|H_1)$. En consecuencia,

$$\begin{aligned} \text{Error Tipo II} &= \Pr(\overline{X}_n < k|H_1) = \Pr(\textit{No rechazar } H_0|H_1 \textit{ verdadera}) \\ &= \text{Area 2} \end{aligned}$$

| | | Estado de la Naturaleza (la verdad) | |
|---|---|---|---|
| | | $H_0$ | $H_1$ |
| Decisión | $H_0$ | O.K. | Error Tipo II |
| | $H_1$ | Error Tipo I | O.K. |

# Test de Hipótesis: Conceptos Generales I

- Significancia Estadística y Poder de un Test:

$$\text{Significancia Estadística} = \Pr(\text{Error Typo I}) = \alpha$$

$$\begin{aligned} \text{Poder} &= \Pr(\textit{Rechazar } H_0 | H_1 \textit{ verdadera}) \\ &= 1 - \Pr(\underbrace{\textit{No Rechazar } H_0 | H_1 \textit{ verdadera}}_{\textit{Error Tipo II}}) \\ &= 1 - \beta \end{aligned}$$

- Estadístico del Test : estadístico usado para aceptar/rechazar $H_0$; se denota $T(\boldsymbol{X}) = \dfrac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$
- Región de Rechazo (o Región Crítica): Valores de $\overline{X}_n$ o $T(\boldsymbol{X})$ que permiten rechazar $H_0$. Often has the form: $T(\boldsymbol{X}) > c$, o $\overline{X}_n > k$)
- Distribución de la Nula: distribución de $T(\boldsymbol{X})$ cuando $H_0$ es verdadera.

# Test de Hipótesis: Conceptos Generales II

- La distribución de la nula es util para calcular el nivel de significancia del test.

$$\begin{aligned}
\Pr(\text{Error Tipo I}) &= \Pr(\overline{X}_n > k | H_0) \\
&= \Pr\left( \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{k - \mu_0}{\sigma/\sqrt{n}} \middle| H_0 \right) \\
&= 1 - \Phi\left( \frac{k - \mu_0}{\sigma/\sqrt{n}} \right)
\end{aligned}$$

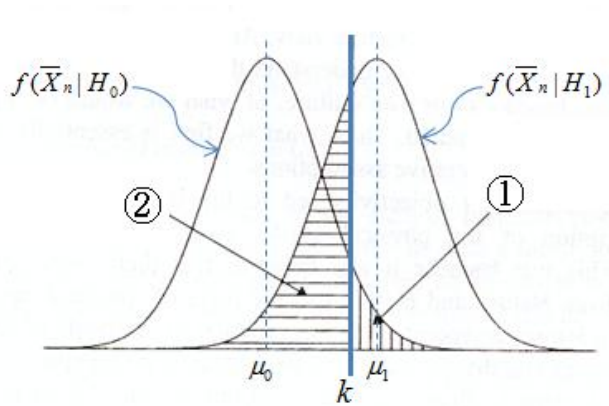
- También podemos calcular la distribución de $T(\boldsymbol{X})$ cuando $H_1$ es verdadera.

$$\begin{aligned}
\Pr(\text{Error Tipo II}) &= \Pr(\overline{X}_n < k | H_1) \\
&= \Phi\left( \frac{k - \mu_1}{\sigma/\sqrt{n}} \right)
\end{aligned}$$

# Test de Hipótesis: Conceptos Generales III

- Nótese que aumentar $k$ implica:
  1. Una reducción en la probabilidad de Error Tipo I (i.e., una reducción en el nivel de significancia del test)
  2. Un aumento en la probabilidad de Error Tipo II (i.e., una reducción en el poder estadístico del test).
- En efecto, cuando elegimos $k$ nos enfrentamos a un trade off con dos tipos de error. **Que podemos hacer para reducir ese trade-off?**

Figure: Probability of two types of errors



# Nivel de Significancia y P-valor

- El resultado del test de hipótesis puede ser rechazar o no rechazar $H_0$.
- El *p-valor* es una medida intermedia para evaluar $H_0$ versus $H_1$:
  - Dada la muestra de datos $\mathbf{x} = x_1 \dots x_n$ , definimos $t^* = T(\boldsymbol{X}) = \dfrac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$ como el test estadístico de interés.
  - El p-valor se define como:

$$\text{p-valor} = \Pr(|t| > |t^*||H_0),$$

  - El p-valor es la probabilidad de observar un valot $t$ mayor a $t^*$ , condicional en que la $H_0$ es verdadera. En otras palabras, el p-valor es el area a la derecha del valor del test estadístico en una distribución normal estándar.
- Note que un p-valor $< \alpha$ implica que el valor del test estadístico es mayor al valor crítico, y por tanto rechazamos $H_0$.
- Una definición equivalente del p-valor es "el minimo nivel de significancia que, dada nuestra muestra, nos llevaría a rechazar la hipótesis nula".

# Nivel de Significancia y P-valor

- Hay 4 pasos claves para usar el p-valor para hacer test de hipótesis:

1. Definir hipótesis nula e hipótesis alternativa.
2. Asumir que la hipótesis nula es verdadera, y calcular el valor del test estadístico $t = T(X) = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$.
3. Conocemos la distribución de $t$ a través de la tabla t. Luego, calculamos $P - Value = P(|T(X)| > |t^*||H_0)$, con $t^*$el valor crítico $t$ para el cual rechazaríamos la hipótesis nula
4. Definimos el nivel de significancia, $\alpha$, la probabilidad de Error Tipo I — 0.01, 0.05, or 0.10. Comparamos el p-valor con $\alpha$. Si $p - valor \leq \alpha$, rechazamos la hipótesis nula en favor de la hipótesis alternativa. Si $p - valor > \alpha$, no rechazamos la hipótesis nula.

# Ejemplo: Testeando poderes extra-sensoriales

De un mazo con 52 cartas, una persona saca 20 cartas aleatoriamente, con reemplazo, y en cada sacada trata de adivinar la pinta de la carta. $X_i$ es un indicador de si la pinta de la carta $i-th$ es correctamente adivinada, y $T(X_1 \dots X_{20}) = \sum_{i=1}^{20} X_i$ es el total de intentos exitosos.

- Definimos la hipótesis nula como$H_0$: "la persona no tiene poderes extra-sensoriales"
- Bajo esta hipótesis, $T(X)$ sigue una distribución Binomial con $n = 20$ y $p = \frac{1}{4}$ (prob. de adivinar pinta – el mazo tiene 4 pintas).

# Ejemplo: Testeando poderes extra-sensoriales

- Si la persona logra adivinar $T(x) = 9$ cartas, el p-valor es:

$$\text{p-valor} = \Pr(Bin(20; 1/4) \geq 9) = 0.041$$

- En consecuencia, para un nivel de significancia del 5% rechazamos la hipótesis nula. Pero no la rechazamos al 1%. Cuál sería el minimo valor de $T$ para el cual rechazaríamos $H_0$ al 1%?
- Con $T = 10$, el p-valor es 1.3%. Con $T = 11$, es 0.4%. En efecto, $T = 11$ es el valor crítico para un nivel de significancia del 1%.
- Nótese que sólo rechazamos la hipótesis nula cuando $T$ es suficientemente grande, *por lo tanto decimos que es un test de hipótesis de una cola*.

# Dualidad entre Test de Hipótesis e Intervalo de Confianza

Considere el siguiente test de hipótesis:

$$\begin{aligned} H_0 : \theta &= \theta_0 \\ H_1 : \theta &\neq \theta_0 \end{aligned}$$

- Un test de hipótesis con un nivel de significancia $\alpha$ define una región de aceptación $A(\theta_0)$ tal que si $T(\boldsymbol{X}) \in A(\theta_0) \Rightarrow$ no rechazamos $H_0$
- Un I.C. especifica una región $C(\boldsymbol{X})$ que contiene el valor verdadero $\theta$ con probabilidad $1 - \alpha$.
- Dado un I.C. para $\theta$, podemos construir una región de aceptación para un nivel de significancia $\alpha$
- Supongamos:

$$A(\theta_0) = \{\boldsymbol{X} \text{ tal que } \theta_0 \in C(\boldsymbol{X})\}, \textit{ luego}$$

$$\begin{aligned} \text{Nivel de Significancia} &= \Pr(\boldsymbol{X} \notin A(\theta_0)|H_0) \\ &= \Pr(\theta_0 \notin C(\boldsymbol{X})|H_0) \\ &= \alpha \end{aligned}$$

# Construyendo un test de hipótesis a partir del I.C.

Considere las siguientes hipótesis nula y alternativa:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned}$$

El siguiente test tiene un nivel de significancia $\alpha$:

1. Obtenemos el I.C. $1 - \alpha\%$ para $\mu$:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

2. Rechazamos la nula si el I.C. no contiene a $\mu_0$

Nótese que este test rechaza la nula cuando $\bar{X}$ es excesivamente grande (positivo) o excesivamente pequeño (negativo) relativo a $\mu_0$, por lo tanto decimos que es un test de dos colas.

# Outline

# Comparando la productividad entre agentes

Table: Resumen de datos del Call Center

| nombre | # obs | Prom | DS | % fallo | semana vacaciones |
|---|---|---|---|---|---|
| Felipe | 1,320 | 2.48 | 0.69 | 7.58% | 11 |
| Camila | 1,320 | 3.54 | 0.53 | 7.05% | 26 |
| Javier | 1,320 | 4.04 | 1.48 | 6.97% | 20 |
| Marta | 1,320 | 2.99 | 1.00 | 7.27% | 27 |
| Violeta | 1,320 | 4.56 | 1.82 | 7.88% | 29 |
| Total | 6,600 | 3.52 | 1.41 | 7.48% | |

- Relativo a Violeta, Camila tiene en promedio llamadas mas cortas y menores tasas de fracaso.
- Es esta diferencia explicada por errrores propios del muestreo aleatorio o debido a diferencias de productividad entre ambos agentes?

# Test de Diferencia de Medias

- Considere dos poblaciones distribuidas normalmente, $X \sim N(\mu_x, \sigma_x^2)$ y $Y \sim N(\mu_y, \sigma_y^2)$.
- Se obtienen dos muestras aleatorias, una de cada población $x_1, \cdots, x_n$ ($n$ obs.) y $y_1, \cdots y_m$ ( $m$ obs.)
- Estamos interesados en el siguiente test de hipótesis.

$$\begin{aligned} H_0 : \mu_x &= \mu_y \Leftrightarrow \mu_x - \mu_y = 0 \\ H_1 : \mu_x &\neq \mu_y \Leftrightarrow \mu_x - \mu_y \neq 0 \end{aligned}$$

# Encontrando la Distribución de la Nula I

- Si $\sigma^2$ es conocido, entonces bajo la hipótesis nula ($H_0$) tenemos que:

$$\overline{X} - \overline{Y} \sim N\left(\underbrace{\mu_x - \mu_y}_{=0}, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

y en efecto,

$$Z = \frac{\overline{X} - \overline{Y}}{\left[\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right]^{\frac{1}{2}}} \sim N(0,1)$$

- En cambio, si $\sigma^2$ fuese desconocida pero común a ambas poblaciones, primero estimamos la varianza de la diferencia considerando ambas muestras:

$$S_p^2 = \frac{S_x^2(n-1) + S_y^2(m-1)}{m+n-2}$$

# Encontrando la Distribución de la Nula II

- Notar que $(m+n-2)S_p^2/\sigma^2 \sim \chi^2_{n+m-2}$, ya que es la suma de dos v.a. independientes que se distribuyen chi-cuadrado.
- Basado en el mismo argumento usado para derivar la distribución t, tenemos que:

$$
\begin{aligned}
t &= \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \\
&= Z/\sqrt{S_p^2/\sigma^2} \\
&= N(0,1)/\sqrt{\chi^2_{n+m-2}/(n+m-2)} \\
&\sim t - \text{student, con d.f.} = n+m-2
\end{aligned}
$$

# Encontrando la Distribución de la Nula III

- Cuando $n$ y $m$ son grandes, el test es valido para cualquier distribución poblacional. Basado en el TLC, el estadístico t sigue una normal estandar.

**Ejercicio: diferencia en la duración de las llamadas**

Testear si Camila tiene duraciones de llamadas menores a las de Camila. muestre el estadístico t, el valor-p y el resultado a un nivel de significancia del 5%.

# Test de una cola vs Test de dos colas

- *Test de dos colas*: cuando el test es diseñado para rechazar valores extremos del estadístico (altos o bajos), i.e., una región de rechazo tal que $|t| > t_{1-\frac{\alpha}{2}}$.
- La región de rechazo también puede ser definida a través de un I.C.

$$(\overline{X} - \overline{Y}) \pm t_{1-\frac{\alpha}{2}} \cdot S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}$$

y rechazar la nula cuando el intervalo no contenga el valor cero.

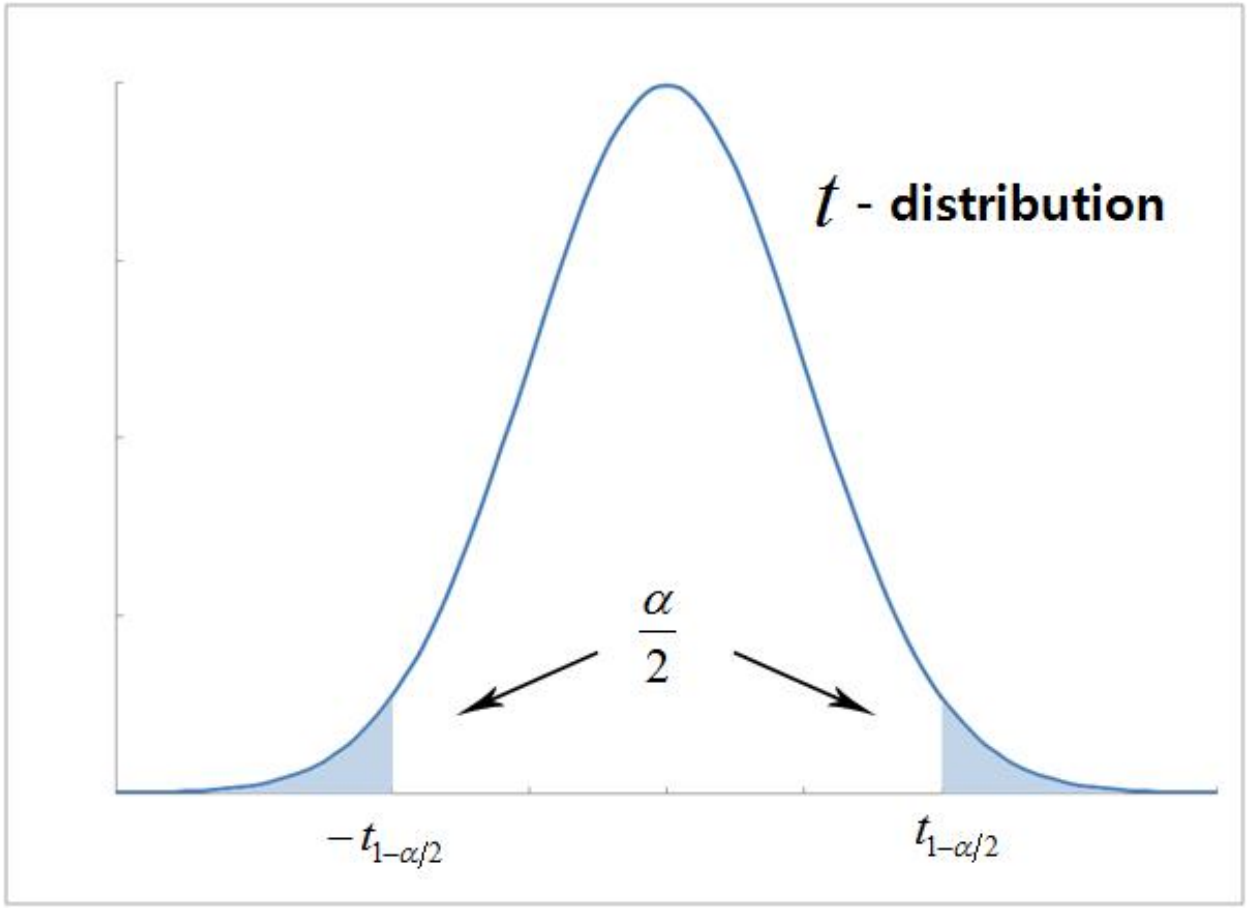
- Si el test sigue el siguiente diseño:

$$\begin{aligned} H_0 : \mu_x &= \mu_y \\ H_1 : \mu_x &> \mu_y \end{aligned}$$

decimos que es un test de una cola, i.e., para un test de tamaño $\alpha$ rechazamos la nula cuando $t > t_{1-\alpha}$.

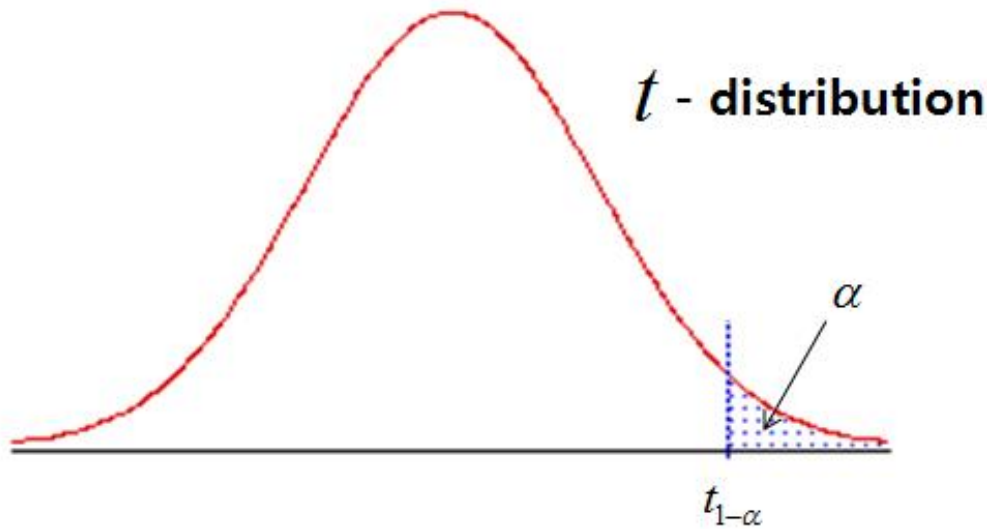
# Test de dos colas

Figure: Valores críticos para un test de dos colas



# Test de una cola

Figure: Valores críticos para un test de una cola



# Testeando diferencias en proporciones I

- Son las tasas de fracaso de Camila y Javier diferentes?
- Esta pregunta puede ser formulada como un test de diferencia de proporciones entre dos poblaciones. Sea $p_i$ la probabilidad de fracaso en la población $i$. Definimos el test como:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

- Sea $x_i$ el numero de fracasos y $\hat{p}_i = \frac{1}{n}\sum_{i=1}^{n} x_i$ la proporción de fracasos en la muestra de la población $i$, con $n$ el tamaño muestral. Sabemos que $\hat{p}_i$ es un estimador insesgado de $p_i$ con un error estándar $SE(\hat{p}_i) = \sqrt{\frac{p_i(1-p_i)}{n_i}}$.
- Dado que las muestras son independientes,

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2).$$

# Testeando diferencias en proporciones II

- Bajo la hipótesis nula ($H_0$), $p_1 = p_2 = p$ y por tanto $Var(\hat{p}_i) = \frac{p(1-p)}{n_i}$.
- Definimos el test estadístico**:**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

  donde $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ es un estimador de $p$ (la tasa de fracaso bajo la hipótesis nula).
- Usamos ambas muestras para mejorar la precisión en la estimación de $p$.
- El error estándar de $\hat{p}$ puede ser estimado reemplazando $(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2})$ dentro de la raíz cuadrada en el denominador.
- **Distribución de la Nula:** Bajo $H_0$, y para un $n_1$ y $n_2$ suficientemente grandes, el test estadístico sigue una normal estándar, $z \sim N(0, 1)$.

# Testeando diferencias en proporciones III

- **Rejection region:** (test de dos colas)
  - Construimos el I.C. $(1-\alpha)$ para la diferencia en proporciones:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

    donde $z_{\alpha/2}$ es el valor crítico obtenido de una normal estándar, tal que $\Pr(N(0,1) > z_{\alpha/2}) = \alpha/2$.
  - Rechazamos $H_0$ si el I.C. no contiene el valor cero.

**Ejercicio**

Testear si la tasa de fracasos de Camila y Javier son estadísticamente distintas. Indicar el test estadístico, el valor-p y el resultado del test para un nivel de significancia de 5%.

# Outline

# Test de diferencia de varianzas

- Considere dos poblaciones distribuidas normalmente: $X \sim N(\mu_x, \sigma_x^2)$ y $Y \sim N(\mu_y, \sigma_y^2)$.
- Dos muestras aleatorias independientes son seleccionadas de cada población, $x_1, \cdots, x_n$ ($n$ obs.) y $y_1, \cdots y_m$ ($m$ obs.)
- Considere el siguiente test de hipótesis:

$$\begin{aligned} H_0 : \sigma_x^2 &= \sigma_y^2 \\ H_1 : \sigma_x^2 &\neq \sigma_y^2 \end{aligned}$$

- Usaremos el test estadístico $S_x^2/S_y^2$, el ratio de las varianzas muestrales, para ejecutar el test.

# Encontrando la distribución de la Nula

- Bajo la hip. nula, $\sigma_x = \sigma_y$, por lo tanto:

$$\frac{S_x^2}{S_y^2} = \frac{\frac{(n-1)S_x^2}{\sigma_x^2} \cdot \frac{1}{(n-1)}}{\frac{(m-1)S_y^2}{\sigma_y^2} \cdot \frac{1}{(m-1)}} \sim \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}$$

**Definition**

Distribución-F

$$\frac{\chi_n^2 \Big/ n}{\chi_m^2 \Big/ m} \sim F_{n,m}$$

donde $\chi_m^2$ y $\chi_n^2$ son v.a. independientes distribuidas chi-cuadrado

# Región de rechazo del test

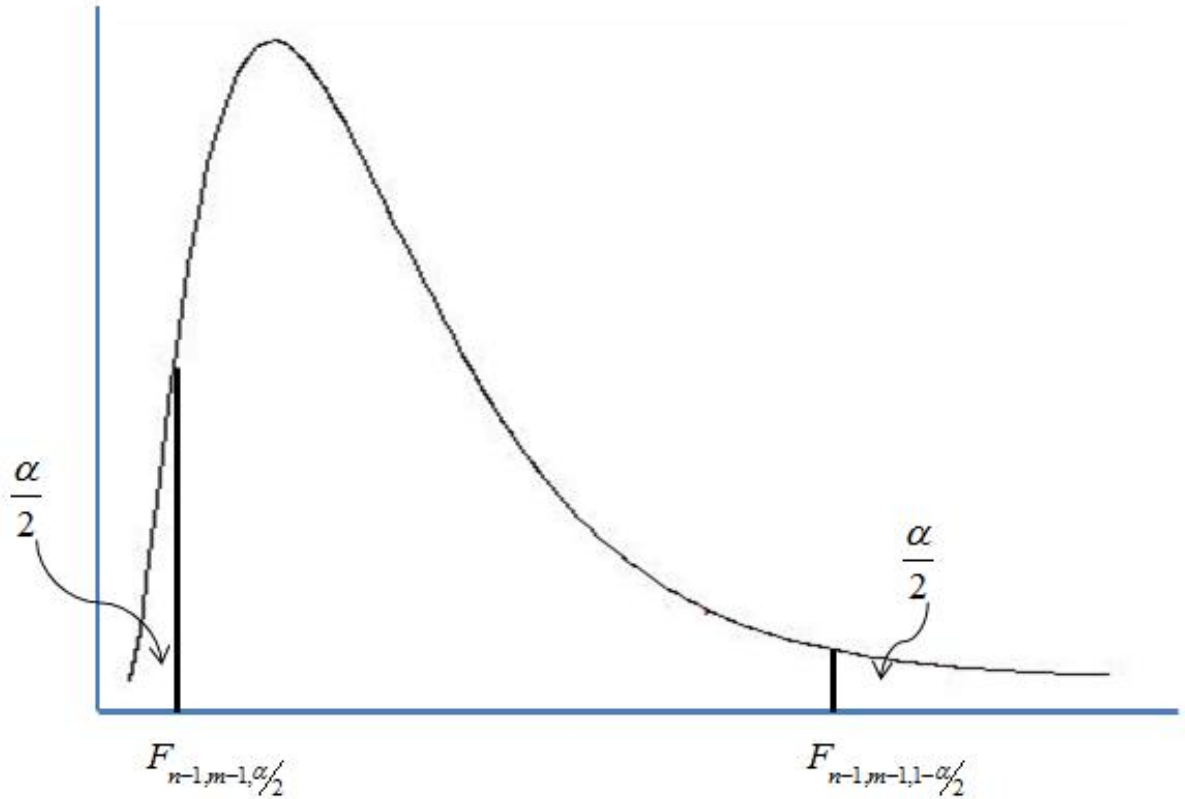
- Intuición: rechazar la nula si el ratio $S_x^2/S_y^2$ está "muy lejos" de 1.
- Para un test de tamaño $\alpha$, rechazar si $S_x^2/S_y^2$ no está contenido en el intervalo:

$$\left[F_{n-1,m-1}\left(\frac{\alpha}{2}\right),\ F_{n-1,m-1}\left(1-\frac{\alpha}{2}\right)\right],$$

  donde los límites del intervalo están dados por el percentil $\alpha/2$ (superior e inferior) de la distribución F
- Este test de dos colas es equivalente al siguiente test de una cola:
  - Construya el estadístico usando la muestra con la varianza mas grande en el numerados, es decir, $S_x^2/S_y^2>1$.
  - Rechazar la nula si $S_x^2/S_y^2 > F_{n-1,m-1}\left(1-\frac{\alpha}{2}\right)$.

# Test-F (dos colas)

Figure: F-distribution



# Conclusion

- Las muestras pueden ser usadas para aprender acerca de características relevantes de la población de interés.
- Conceptos básicos: estimador insesgado, estimador consistente, intervalos de confianza, test de hipótesis.
- Existen diferentes estrategias para estimar parametros poblaciones
  - Métodos exactos: requieren supuestos fuertes respecto a la distribución poblacional desde la cual proviene la muestra.
  - Métodos asintóticos: descansan sobre el Teorema del Límite Central, pero sólo aplican para muestras grandes.

# Appendix

# Outline

# Sampling distribution with a normal population

- We assume that $X \sim \text{Normal}(\mu, \sigma^2)$ where $\mu$ and $\sigma^2$ are unknown.
- We show how to construct a confidence interval for $\mu$ using the following 3 results:
  - Result 1: $\overline{X}$ and $S^2$ are independent random variables.
  - Result 2: The statistic $\frac{(n-1)S^2}{\sigma^2}$ follows a *chi-square* distribution with $n-1$ degrees of freedom.
  - Result 3: The statistical distribution $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ depends only on $n$ and follows a *t-student* distribution with $n-1$ degrees of freedom.

# Derivation of Result 1

We can write the sample variance as:

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \\
&= \frac{1}{n-1}\left\{(X_1-\overline{X})^2+\sum_{i=2}^{n}(X_i-\overline{X})^2\right\} \\
&= \frac{1}{n-1}\left\{\left[(X_1-\overline{X})-\sum_{i=1}^{n}(X_i-\overline{X})\right]^2+\sum_{i=2}^{n}(X_i-\overline{X})^2\right\} \\
&= \frac{1}{n-1}\left\{\left[\sum_{i=2}^{n}(X_i-\overline{X})\right]^2+\sum_{i=2}^{n}(X_i-\overline{X})^2\right\}
\end{aligned}
$$

This is a function of $(X_i-\overline{X})$, $i=2,\cdots,n$, only. If we show that the vector is independent of $\overline{X}$, we're done.

# Derivation of Result1(ii)

Use the following monotonic transformation:

$$\begin{aligned} y_1 &= \overline{x} \\ y_2 &= x_2 - \overline{x} \\ &\cdots \\ y_n &= x_n - \overline{x} \end{aligned}$$

...with its inverse given by

$$\begin{aligned} x_2 &= y_1 + y_2 \\ x_3 &= y_1 + y_3 \\ &\cdots \\ x_n &= y_1 + y_n \\ x_1 &= \sum_{i=1}^{n} x_i - \sum_{i=2}^{n} x_i \\ &= n\overline{x} - \sum_{i=2}^{n} x_i \\ &= ny_1 - \left[ (n-1)y_1 + \sum_{i=2}^{n} y_i \right] \\ &= y_1 - \sum_{i=2}^{n} y_i \end{aligned}$$

# Derivation of Result 1 (iii)

The joint distribution of $X_1, X_2 \cdots, X_n$ is

$$\begin{aligned} f(x_1, \cdots, x_n) &= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\left[x_1^2 + \sum_{i=2}^{n} x_i^2\right]} \end{aligned}$$

Therefore, the joint distribution of $Y_1, Y_2 \cdots, Y_n$ is

$$\begin{aligned} f(y_1, \cdots, y_n) &= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\left\{\left[y_1 - \sum_{i=2}^{n} y_i\right]^2 + \sum_{i=2}^{n}(y_1 + y_i)^2\right\}} \cdot |J| \\ &= \frac{|J|}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(ny_1)^2} e^{-\frac{1}{2}\left\{\left[\sum_{i=2}^{n} y_i\right]^2 + \sum_{i=2}^{n} y_i^2\right\}} \end{aligned}$$

$\Rightarrow y_1$ is independent of $(y_2, \cdots, y_n)$.
This proves Step 1.

# Derivation of Result 2 (i)

We use the following results:

1. If $Z \sim N(0,1)$,then $Z^2$ follows a chi-square distribution with $df = 1$ ($df$ =degrees of freedom), that is,

$$Z^2 \sim \chi^2_1$$

The density function of $Y := Z^2$ is given by,

$$f(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}, \qquad 0 < y < \infty.$$

2. If $X_1, X_2 \cdots, X_n$ are independent and $X_i \sim \chi^2_{p_i}$(chi-square distribution with $p_i$ degrees of freedom), then

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n \sim \chi^2_{p_1+p_2+\cdots+p_n}.$$

That is, the sum of r.v with chi-square distribution follows a chi-square distribution with $df = \sum df_i$ .

# Derivation of Result 2 (ii)

- These two results imply that if $Z_1, Z_2 \cdots, Z_n$ are independent r.v normally distributed as standard normal, then $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$
- This implies that $W = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi_n^2$. The moment generating function (m.g.f.) of $W$ is $M_W(t) = (1-2t)^{-\frac{n}{2}}$.
- This r.v can be write as:

$$
\begin{aligned}
W &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X} + \overline{X} - \mu)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 + n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 + \frac{2}{\sigma^2}(\overline{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} \\
&= \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{U} + \underbrace{\left(\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}\right)^2}_{V} \\
W &= U + V
\end{aligned}
$$

# Derivation of Result 2 (iii)

Step 1 implies that $S^2$ and $\overline{X}$ are independent. Therefore $U$ and $V$ are also independent r.v. Therefore, we can factor the moment generating function as:

$$M_W(t) = M_U(t) \cdot M_V(t)$$

The m.g.f of the r.v $U$ can be get as:

$$\begin{aligned} M_U(t) &= \frac{M_W(t)}{M_V(t)} \\ &= \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} \\ &= (1-2t)^{-\frac{(n-1)}{2}} \\ &\Rightarrow \text{f.g.m.de } \chi^2_{\text{n}-1} \end{aligned}$$

Therefore, follows that $U \sim \chi^2_{n-1}$.

# Derivation of Result 3

Remember that our aim is to find the distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$. This statistic can be write as:

$$t = \underbrace{\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)}_{\sim N(0,1)} \Bigg/ \underbrace{\left(\frac{(n-1)S^2}{\sigma^2} \Big/ (n-1)\right)^{\frac{1}{2}}}_{\sim \sqrt{\chi^2_{n-1}/(n-1)}}$$

## Definition

[t-student distribution]

If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ are independent, then $\frac{Z}{\sqrt{U/n}}$ has a *t-student* distribution. It can be show that the density of this r.v is:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \qquad -\infty < t < \infty.$$

# References


- William H. Greene, Econometric Analysis, Editorial Pearson, 7th Edition, Appendix B and C
- John A. Rice, Mathematical Statistics and Data Analysis, Editorial Thomson Brooks/Cole, Third Edition. Chapters 6 y 7.
- Jeffrey Wooldridge, Introductory Econometrics: A Modern Approach, Editorial South-Western Cengage Learning, Appendix B and C.