Intervalos de confianza
Muestras pequeñas

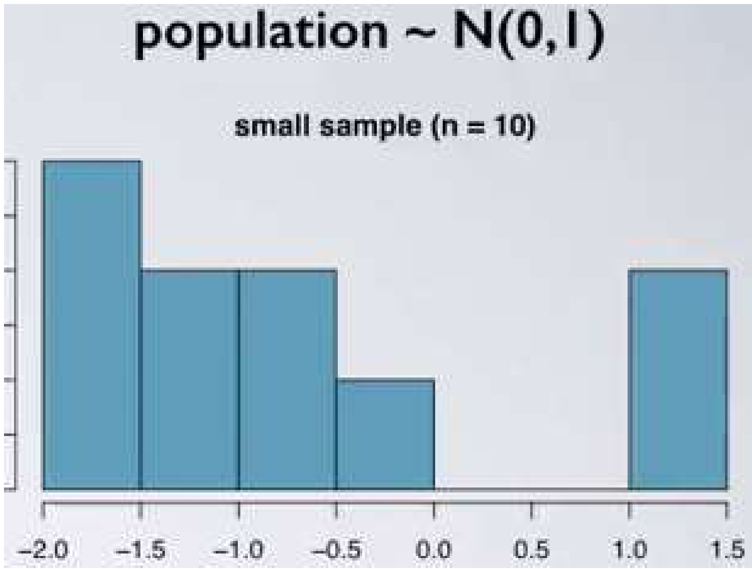
¿Qué ocurre cuando n<30?

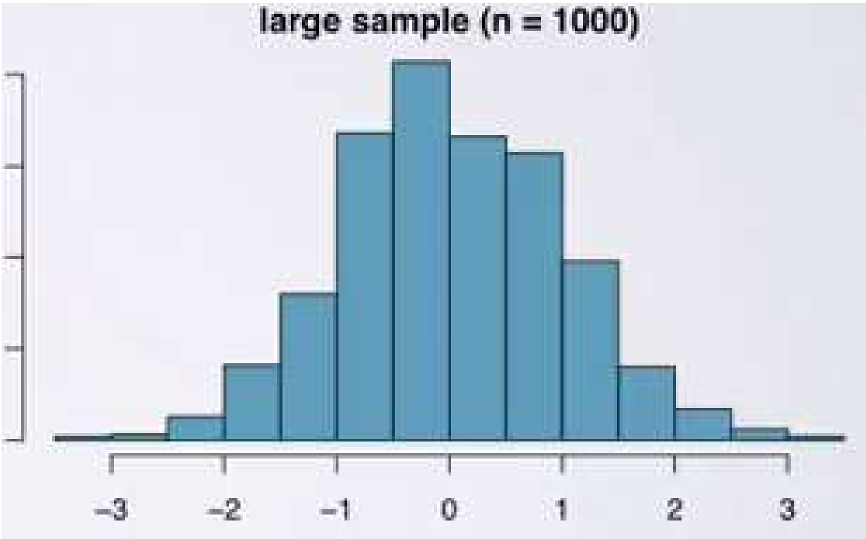
- No siempre se tiene la posibilidad de contar con una muestra grande.
- Vamos a focalizarnos en **muestras pequeñas** cuando el estadístico es la **media muestral**
- Recordemos porque necesitamos una **muestra grande**
- Siempre y cuando las observaciones sean i.i.d., y la distribución poblacional no demasiado asimétrica, una muestra grande nos aseguraba que
  - La distribución muestral de la media se aproxima a la normal a medida que n (tamaño de la muestra) crecía
  - Y el estimador del error estándar: $SE = \frac{s}{\sqrt{n}}$ es confiable donde s es el estimador de σ , el desvío poblacional que por lo general es desconocido.

¿Qué ocurre cuando n<30?

- El TCL asevera que la distribución muestral de $\overline{X}$ es aprox. Normal cualquiera sea el tamaño de la muestra siempre y cuando la distribución poblacional sea aprox. normal .

- Sin embargo, no es fácil de verificar en muestras pequeñas la condición de normalidad. Ambas muestras (n=10 y n=1000) provienen de una N(0,1)





- Es difícil determinar a partir de una muestra pequeña cual es la distribución de la que provienen.

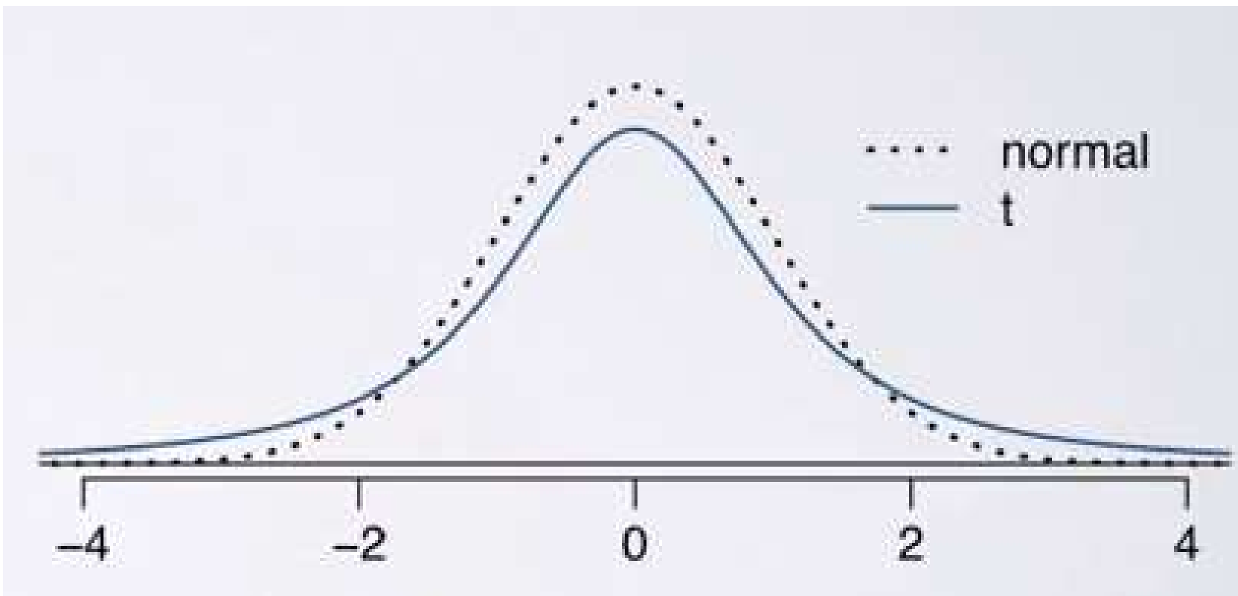
¿Qué ocurre cuando n es pequeño?

- Vimos que si **n es grande**, y σ es desconocido, estimamos σ con s
- Sin embargo, cuando n es pequeño y σ es desconocido (casi siempre), también podemos utilizar s como el estimador natural de σ, pero el hecho de que n sea pequeño torna a s menos confiable.
- Para mitigar esta mayor incertidumbre en s y continuar reteniendo la confianza del 95% en la construcción de los intervalos de confianza, deberíamos entonces aumentar el ancho del intervalo.
- Luego deberíamos trabajar con una distribución que de cuenta de la necesidad de un intervalo más ancho.
- Por lo tanto, la distribución normal estandarizada Z es reemplazada por la t de Student.

La distribución t de Student

- La t de Student también es simétrica alrededor de la media=0, con forma de campana, pero con colas más pesadas, i.e. es más probable tener más observaciones más allá de 2 desvíos estándar respecto de la media si se la compara con la distr. Normal estándar.

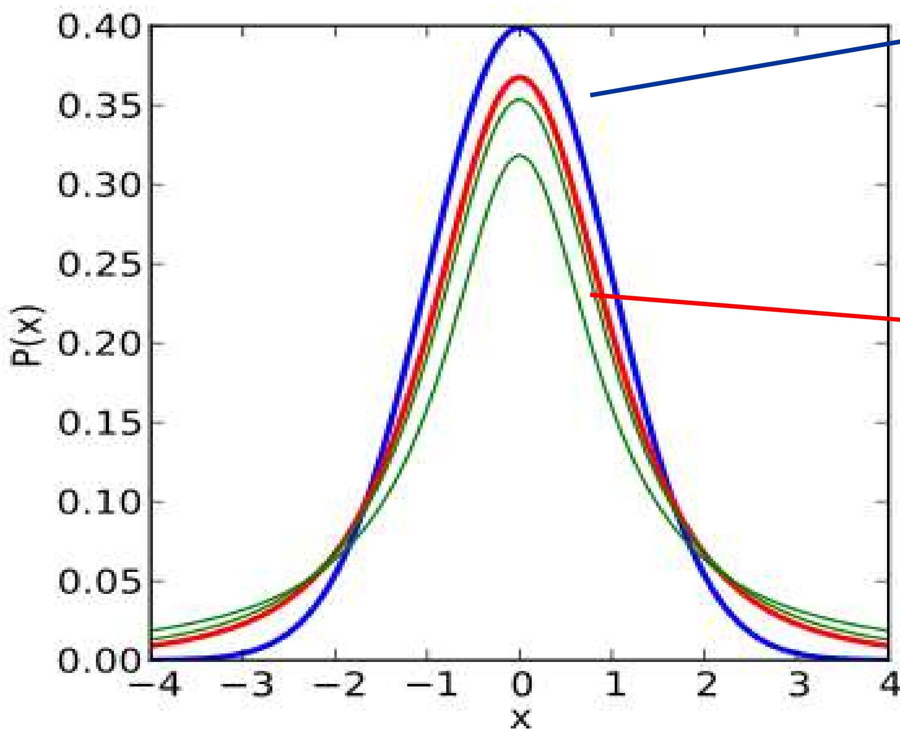
- Las colas más pesadas son las que van a mitigar la mayor incertidumbre originada en el cálculo del SE



La distribución t de Student

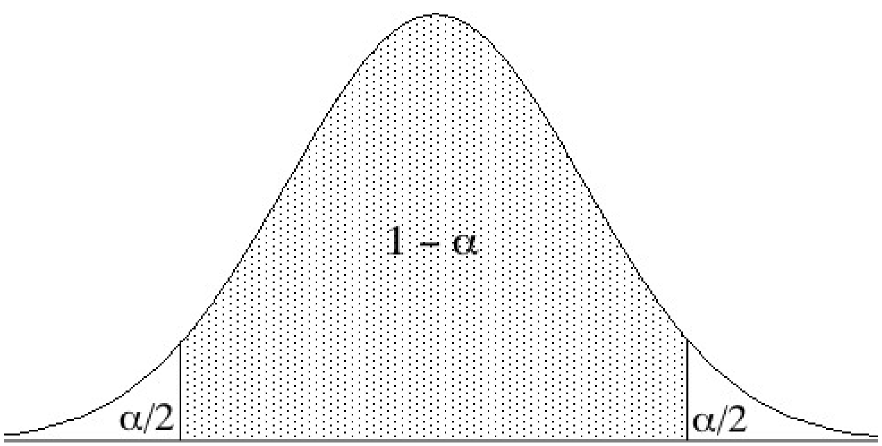
- La t de Student tiene un solo parámetro, llamado **grados de libertad**, que determina cuan pesadas son las colas de la distribución.

- ¿Qué ocurre con la forma de la distribución cuando los grados de libertad se incrementan?



¿Cuándo y cómo se utiliza la t de Student ?

- Cuando deseamos construir un intervalo de confianza para la media y el
  - σ es desconocido
  - Tamaño de la muestra pequeño
- El intervalo de confianza (región de confianza=1-α) se calcula de la misma manera pero en lugar de Z utilizamos t



$$\bar{X} \pm t_{n-1,\alpha/2} * \frac{S}{\sqrt{n}}$$

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

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

Comparando la N(0,1) y la t

- Calcular

a. $P\left(Z \geq z_{0.025}\right) = 0.025$

b. $P\left(t \geq t_{50,0.025}\right) = 0.025$

c. $P\left(t \geq t_{10,0.025}\right) = 0.025$

El origen de la t de Student

- William Gosset, 1876-1937. Estudió química y matemática.
- Se unió a la cervecería Guinness, donde llegó a ocupar la posición más alta en el área de investigaciones de la compañía.
- Su preocupación era estudiar los tipos de cebadas para mejorar la calidad de la cerveza. Rara vez disponía de muestras grandes. Eso lo llevó a estudiar una distribución para muestras pequeñas.
- Llega a un acuerdo con la compañía para publicar sus trabajos estadísticos bajo el seudónimo de **Student**.

Ejemplo

- Cierta empresa está implementado un programa de adiestramiento por computadora para sus empleados. La empresa decide adiestrar a 15 empleados. La tabla muestra los tiempos de adiestramiento.

Tiempo de adiestramiento en dias							
Empleado	Tiempo		Empleado	Tiempo		Empleado	Tiempo
1	52		6	59		11	54
2	44		7	50		12	58
3	55		8	54		13	60
4	44		9	62		14	62
5	45		10	46		15	63

- Calcular el intervalo de confianza al 95% para la media poblacional

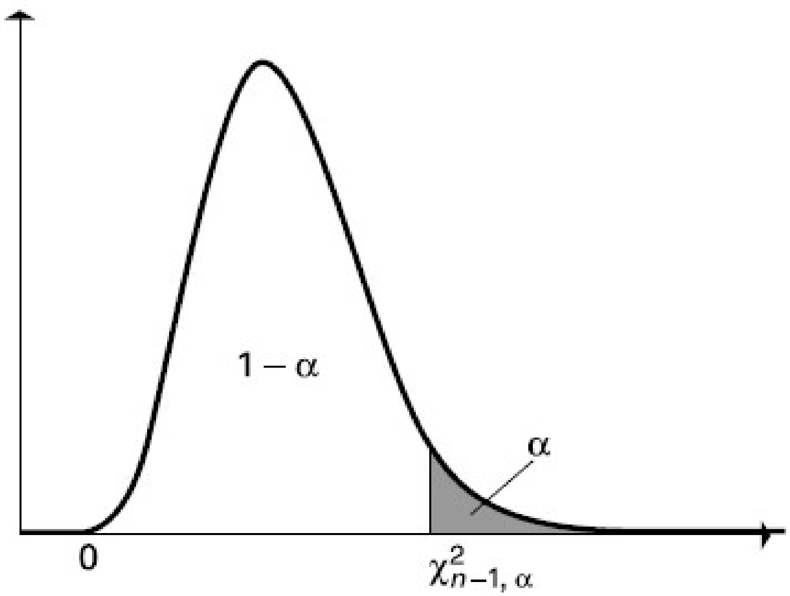
Intervalos de confianza de la varianza de una distribución normal

- Supongamos que de una población que sigue una distribución normal con varianza σ^2 se extrae una muestra aleatoria de **n observaciones** cuya varianza es s^2 . Entonces

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

- Sigue una distribución χ^2_{n-1} con n-1 grados de libertad
- Recordemos que

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



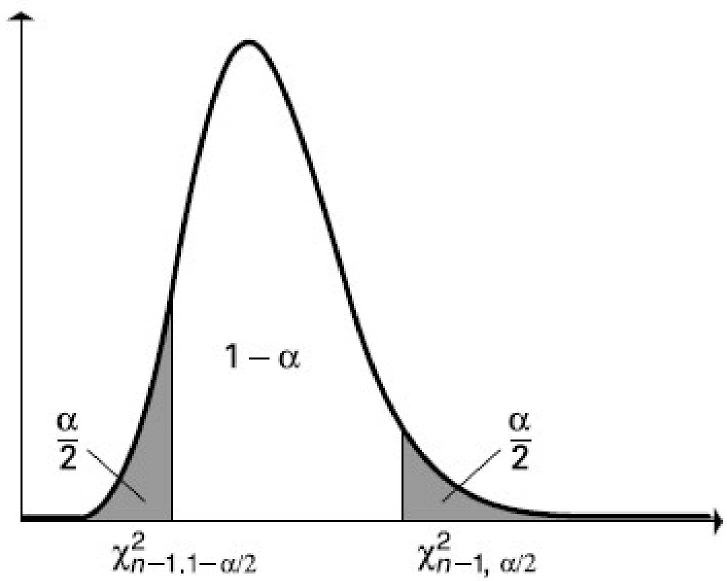
Intervalos de confianza de la varianza de una distribución normal

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

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

- Por lo tanto

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



- Finalmente

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

Intervalos de confianza de la varianza de una distribución normal

- Ejemplo: Supongamos que queremos hallar un par de números tal que la probabilidad de que una variable aleatoria chi-cuadrado con 8 grados de libertad se encuentre entre estos números es 0.90.

Intervalos de confianza de la varianza de una población normal

- Supongamos que hay una muestra aleatoria de n observaciones extraídas de una población que sigue una distribución normal de varianza σ^2. Si la varianza muestral observada es s^2, entonces un intervalo de confianza al $100(1-\alpha)\%$ de la varianza poblacional es:

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

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

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

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

Intervalos de confianza de la varianza de una distribución normal

- El director de Aceros Norte, quiere evaluar la variación de la temperatura en el nuevo horno eléctrico de la empresa. Se obtiene una muestra aleatoria de 25 temperaturas durante 1 semana y se observa que la varianza muestral es $s^2 = 100$. Halle el intervalo de confianza al 95% de la varianza poblacional de la temperatura.

- Supuestos???

Resumiendo...

Parámetro	Muestra	Distribución poblacional	σ	Intervalo de confianza
Media	n grande $(n \geq 30\,?)$	Cualquiera	Conocida	$\bar{X} \pm Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
Media	n grande $(n \geq 30\,?)$	Cualquiera	Desconocida	$\bar{X} \pm Z_{\alpha/2} \dfrac{S}{\sqrt{n}}$
Media	$n \leq 30$	Debe ser aprox. normal	Conocida	$\bar{X} \pm Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
Media	$n \leq 30$	Debe ser aprox. normal	Desconocida	$\bar{X} \pm t_{n-1,\alpha/2} \dfrac{S}{\sqrt{n}}$
Proporción	$np \geq 10$ $n(1-p) \geq 10$			$\hat{p} \pm Z_{\alpha/2} \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
Varianza		Normal		$\dfrac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \dfrac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$