Statistics: Confidence Intervals

- How can we estimate the mean from data and quantify the uncertainty in our estimate?
- Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ. A confidence interval $[A, B]$ for the mean μ with confidence level $1-\alpha$ satisfies $\mathbb{P}[A \le \mu \le B] = 1-\alpha$ where A and B are functions of $X_1, \dots, X_n$.
- If we estimate the mean with the sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, we get a confidence interval $[M_n - \epsilon, M_n + \epsilon]$ where we need to properly select $\epsilon > 0$ to get confidence level $1-\alpha$.

$\rightarrow$ $\mathbb{P}[M_n - \epsilon \le \mu \le M_n + \epsilon]$

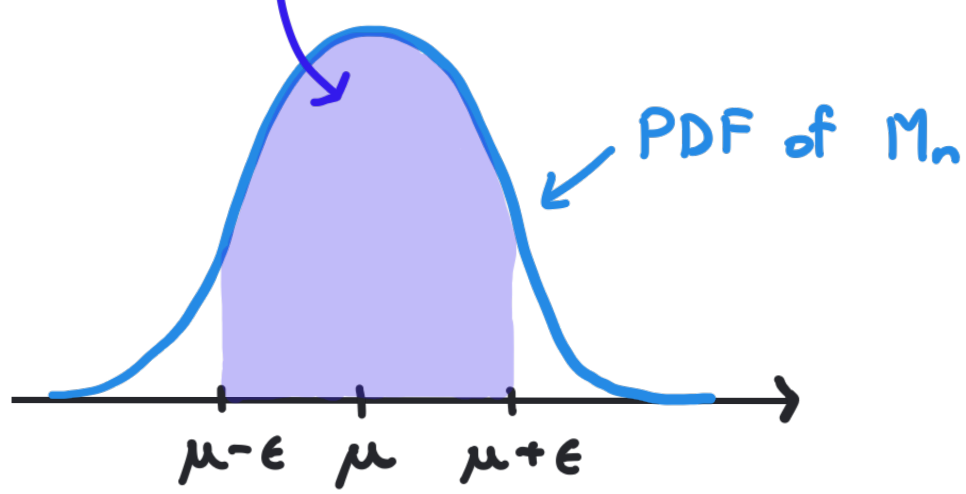
$= \mathbb{P}[\mu - \epsilon \le M_n \le \mu + \epsilon]$

Subtract $M_n + \mu$ from all sides, multiply by -1

$\rightarrow$ $\mathbb{E}[M_n] = \mu$

$\rightarrow$ Approximate M_n as Gaussian based on Central Limit Theorem.

Select ϵ to make this area $= 1-\alpha$.



• Confidence Interval for the Mean: Known Variance

→ Given i.i.d. $X_1, \ldots, X_n$ with known variance σ^2, determine a confidence interval with confidence level $1-\alpha$.

① Calculate the sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

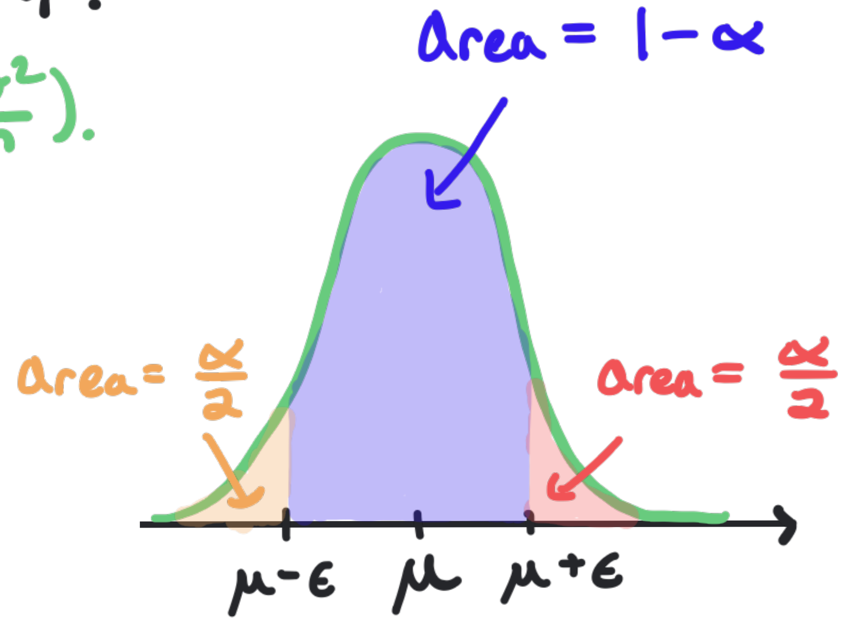
Assume M_n is (approximately) Gaussian$(\mu, \frac{\sigma^2}{n})$.

② Choose $\epsilon > 0$ so that

$$1-\alpha = \mathbb{P}[\mu-\epsilon \le M_n \le \mu+\epsilon]$$

$$= 1 - \left(\boxed{\mathbb{P}[M_n < \mu-\epsilon]} + \boxed{\mathbb{P}[M_n > \mu+\epsilon]} \right)$$

Set to $\alpha/2$. Set to $\alpha/2$.



$$\mathbb{P}[M_n < \mu-\epsilon] \approx \Phi\left(\frac{\mu-\epsilon-\mu}{\sqrt{\sigma^2/n}}\right) = \Phi\left(-\frac{\epsilon\sqrt{n}}{\sigma}\right) = Q\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)$$ Standard Complementary CDF $Q(z)$.

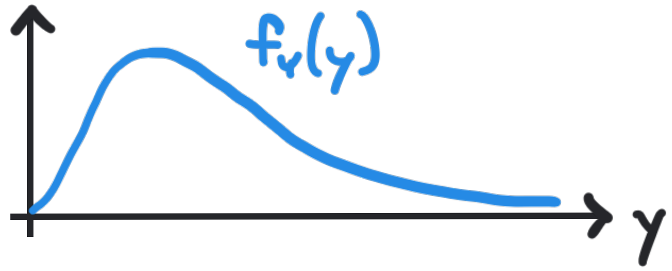
$$\mathbb{P}[M_n > \mu+\epsilon] \approx Q\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)$$ by symmetry $\Rightarrow Q^{-1}\left(\frac{\alpha}{2}\right) = \frac{\epsilon\sqrt{n}}{\sigma}$

③ Overall, $[M_n - \epsilon, M_n + \epsilon]$ with $\epsilon = \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)$ is a confidence interval for the mean with confidence level $1-\alpha$.

MATLAB: $Q^{-1}(z)$ = qfuncinv(z)

- If variance is unknown, estimate using the sample variance.
- We need two new families of random variables.
- If $Z_1, \ldots, Z_n$ are i.i.d. Gaussian(0,1), then $Y = \sum_{i=1}^{n} Z_i^2$ is a chi-squared random variable with n degrees-of-freedom.

→ Mean: n → Variance: $2n$ → PDF Sketch:



→ Shorthand Notation: $Y \sim \chi_n^2$

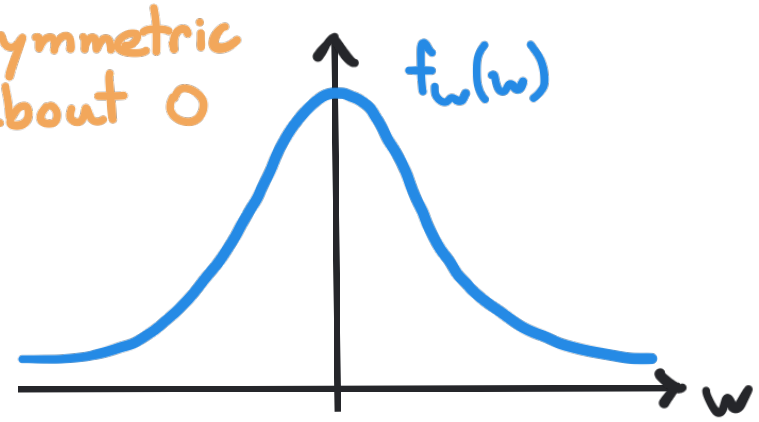
→ CDF: $F_{\chi_n^2}(y)$ evaluate using lookup table or software

- If Z is Gaussian(0,1), $Y \sim \chi_n^2$, and Y and Z are independent, then $W = Z\sqrt{\frac{n}{Y}}$ has a Student's t-distribution with n degrees-of-freedom.

→ Mean: 0 → Variance: $\frac{n}{n-2}$ for $n \geq 3$ → PDF Sketch:

(∞ for $n = 1, 2$)

Symmetric about 0



→ Shorthand Notation: $W \sim T_n$

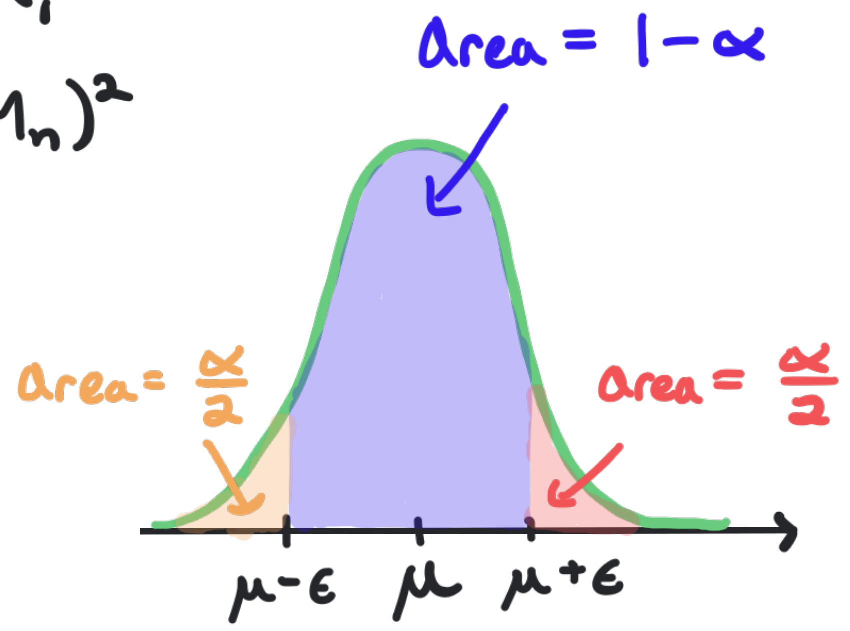
→ CDF: $F_{T_n}(y)$ evaluate using lookup table or software. Converges to $\Phi(y)$ as $n \to \infty$.

• Confidence Interval for the Mean: Unknown Variance

→ Given i.i.d. $X_1, \dots, X_n$ with unknown variance, determine a confidence interval with confidence level $1-\alpha$.

① Calculate the sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

and the sample variance $V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$



② Choose $\epsilon > 0$ so that

$$1-\alpha = \mathbb{P}[\mu-\epsilon \le M_n \le \mu+\epsilon]$$
$$= 1 - \left(\underbrace{\boxed{\mathbb{P}[M_n < \mu-\epsilon]}}_{\text{Set to } \alpha/2.} + \underbrace{\boxed{\mathbb{P}[M_n > \mu+\epsilon]}}_{\text{Set to } \alpha/2.} \right)$$

→ Can argue (see lecture notes) that $\frac{\sqrt{n}(M_n-\mu)}{\sqrt{V_n}}$ has a Student's t-distribution with $n-1$ degrees-of-freedom if $X_1,\dots,X_n$ i.i.d. Gaussian.

$$\mathbb{P}[M_n < \mu-\epsilon] = \mathbb{P}\left[\frac{\sqrt{n}(M_n-\mu)}{\sqrt{V_n}} < \frac{\sqrt{n}(\cancel{\mu}-\epsilon-\cancel{\mu})}{\sqrt{V_n}}\right] = F_{T_{n-1}}\left(-\frac{\epsilon\sqrt{n}}{\sqrt{V_n}}\right) = \frac{\alpha}{2}$$

$\mathbb{P}[M_n > \mu+\epsilon] = \frac{\alpha}{2}$ follows by symmetry

③ $\boxed{\text{Overall, } [M_n-\epsilon, M_n+\epsilon] \text{ where } \epsilon = -\frac{\sqrt{V_n}}{\sqrt{n}} F^{-1}_{T_{n-1}}\left(\frac{\alpha}{2}\right) \text{ is a confidence interval for the mean with confidence level } 1-\alpha.}$

MATLAB: $F^{-1}_{T_{n-1}}(z) = \text{tinv}(z, n-1)$

• Confidence Interval for the Variance:

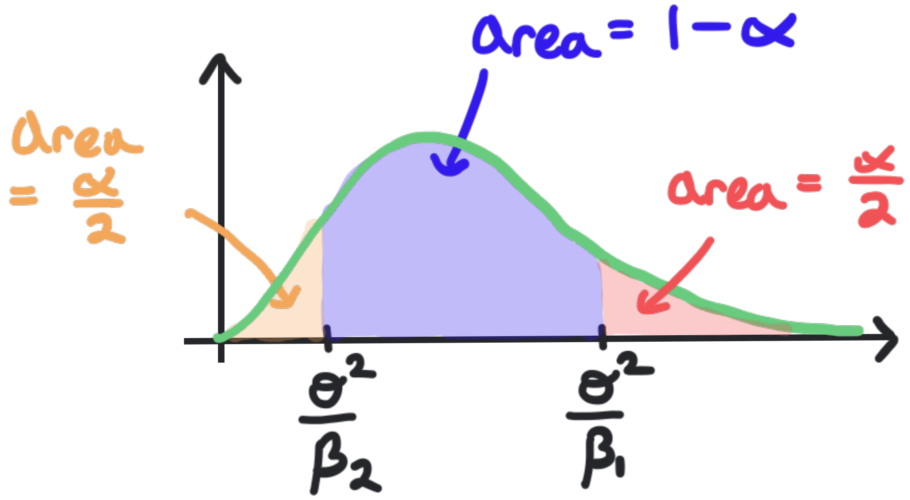
→ For i.i.d. $X_1, \ldots, X_n$, find a confidence interval for the variance $\sigma^2 = \mathrm{Var}[X]$ with confidence level $1-\alpha$.

① Calculate the sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
and the sample variance $V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$

② Pick $0 < \beta_1 < \beta_2$ so that

$$1-\alpha = \mathbb{P}[\beta_1 V_n \le \sigma^2 \le \beta_2 V_n]$$

$$= 1 - \Big(\underbrace{\boxed{\mathbb{P}\Big[V_n < \frac{\sigma^2}{\beta_2}\Big]}}_{\text{Set to } \alpha/2} + \underbrace{\boxed{\mathbb{P}\Big[V_n > \frac{\sigma^2}{\beta_1}\Big]}}_{\text{Set to } \alpha/2} \Big)$$



→ Can argue (see lecture notes) that $\frac{n-1}{\sigma^2} V_n$ has a χ^2-distribution with $n-1$ degrees-of-freedom if $X_1, \ldots, X_n$ i.i.d. Gaussian.

$$\mathbb{P}\Big[V_n < \frac{\sigma^2}{\beta_2}\Big] = \mathbb{P}\Big[\frac{n-1}{\sigma^2} V_n < \frac{n-1}{\cancel{\sigma^2}} \frac{\cancel{\sigma^2}}{\beta_2}\Big] = F_{\chi^2_{n-1}}\Big(\frac{n-1}{\beta_2}\Big) = \frac{\alpha}{2}$$

$$\mathbb{P}\Big[V_n > \frac{\sigma^2}{\beta_1}\Big] = 1 - \mathbb{P}\Big[V_n \le \frac{\sigma^2}{\beta_1}\Big] = 1 - F_{\chi^2_{n-1}}\Big(\frac{n-1}{\beta_1}\Big) = \frac{\alpha}{2} \Rightarrow F_{\chi^2_{n-1}}\Big(\frac{n-1}{\beta_1}\Big) = 1 - \frac{\alpha}{2}$$

③ Overall, $[\beta_1 V_n, \beta_2 V_n]$ where $\beta_1 = (n-1)/F^{-1}_{\chi^2_{n-1}}(1-\frac{\alpha}{2})$, $\beta_2 = (n-1)/F^{-1}_{\chi^2_{n-1}}(\frac{\alpha}{2})$ is a confidence interval for the variance with confidence level $1-\alpha$.

MATLAB: $F^{-1}_{\chi^2_{n-1}}(z)$ = chi2inv(z, n-1)