CSC380: Principles of Data Science

Statistics 4

Credit:
- Jason Pacheco,
- Kwang-Sung Jun,
- Chicheng Zhang
- Xinchen yu

Review: Sample Mean for Bernoulli

Sample mean: $\hat{p} = \frac{1}{N}\sum_i X_i$

Expectation:
$$\mathbf{E}[\hat{p}(X)] = \mathbf{E}\left[\frac{1}{N}\sum_i X_i\right] \overset{(a)}{=} \frac{1}{N}\sum_i \mathbf{E}[X_i] \overset{(b)}{=} \frac{1}{N}Np = p$$

Variance:
$$\mathbf{Var}(\hat{p}) = \mathbf{Var}\left(\frac{1}{N}\sum_i X_i\right) \overset{(a)}{=} \frac{1}{N^2}\mathbf{Var}\left(\sum_i X_i\right) \overset{(b)}{=} \frac{1}{N^2}\sum_i \mathbf{Var}(X_i) \overset{(c)}{=} \frac{1}{N^2}\sum_i p(1-p) = \frac{1}{N}p(1-p) = \frac{1}{N}\mathbf{Var}(X)$$

Review: Sample Mean for Gaussian

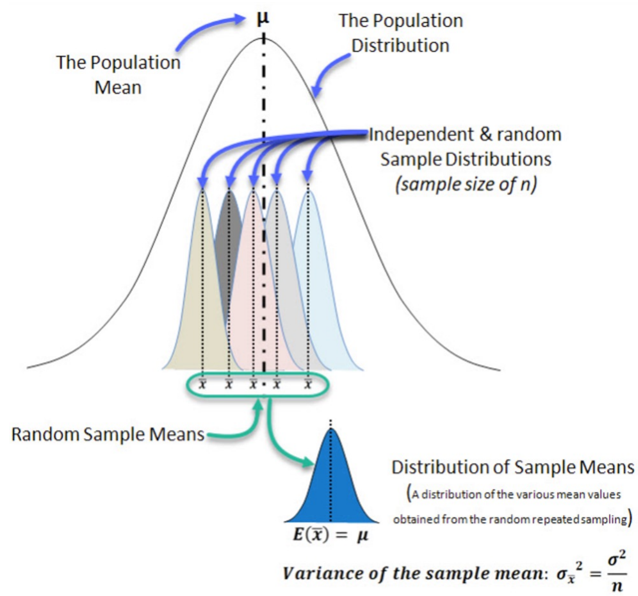
(Property of Gaussian: $\mathrm{E}[X] = \mu_x$, $Var[X] = \sigma_x^2$)

Expectation:

$$\mathbf{E}[\hat{p}(X)] = \mathbf{E}\left[\frac{1}{N}\sum_i X_i\right] \overset{\text{(a)}}{=} \frac{1}{N}\sum_i \mathbf{E}\left[X_i\right]$$

Variance:

$$\mathbf{Var}(\hat{p}) = \mathbf{Var}\left(\frac{1}{N}\sum_i X_i\right) \overset{(a)}{=} \frac{1}{N^2}\mathbf{Var}\left(\sum_i X_i\right) \overset{(b)}{=} \frac{1}{N^2}\sum_i \mathbf{Var}\left(X_i\right) = \frac{1}{N}\mathbf{Var}(X)$$



3

Review: Sample Variance

Sample variance: $\hat{\sigma}^2 = \frac{1}{N}\sum_i (X_i - \hat{\mu})^2$ ← Source of bias: plug-in mean estimate

Expectation: $\mathbf{E}[\hat{\sigma}^2] = \frac{1}{N}\sum_i \mathbf{E}\left[(X_i - \hat{\mu})^2\right] = \text{boring algebra} = \frac{N-1}{N}\sigma^2$

Correcting bias : $\hat{\sigma}^2_{\textbf{unbiased}} = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1}\sum_i (X_i - \hat{\mu})^2$

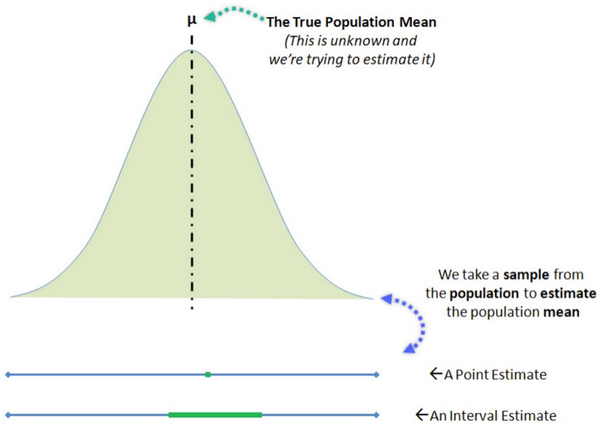
$$E\left[\hat{\sigma}^2_{unbiased}\right] = \sigma^2$$

Biased version has lower MSE: Bias-Variance tradeoff

4

Point estimate vs Interval estimate

5



- **Point estimate**: a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter.

- **Interval estimate**: a range of values constructed from sample data so that the population parameter will likely occur within the range at a specified probability.

5

Confidence Intervals

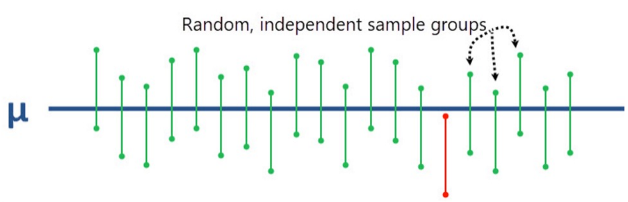
6

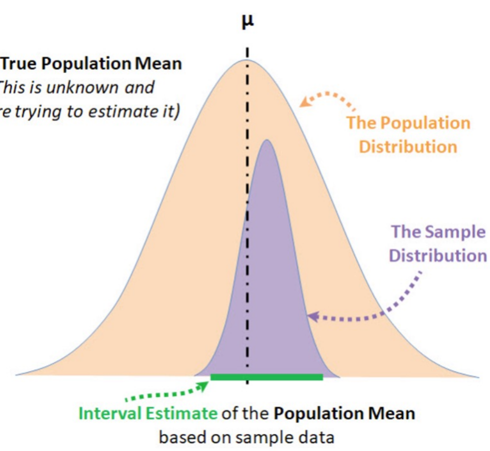
Informally, find an interval such that we are *pretty sure* it encompasses the true parameter value.

Given data $X_1, \ldots, X_n$ and ~~confidence~~ failure rate $\alpha \in (0,1)$ find interval (a,b) such that,

$$P(\theta \in (a,b)) \geq 1-\alpha$$

The interval (a,b) contains the true parameter value θ with probability **at least** $1-\alpha$





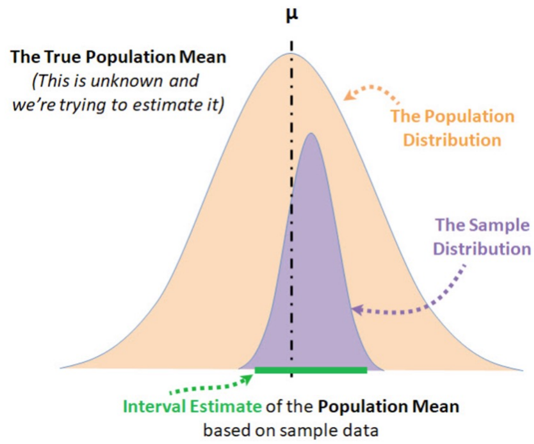
6

Confidence Intervals

7

The interval (a, b) contains the true parameter value θ with probability **at least** $1-\alpha$

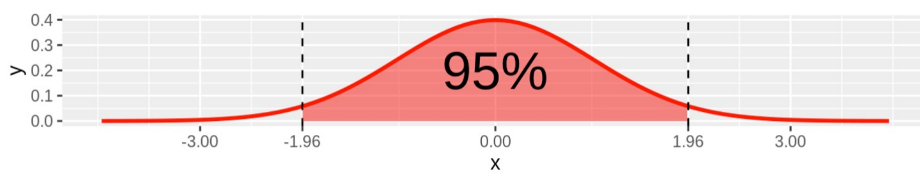
- Intervals must be computed from data: i.e., $a(X_1, \ldots, X_n)$ and $b(X_1, \ldots, X_n)$
- Interval (a,b) is **random**
- parameter θ is **not random** (it is fixed)



- Usually, you compute an estimator $\hat{\theta}$ and then set $a = \hat{\theta} - \epsilon_a$ and $b = \hat{\theta} - \epsilon_b$ for a carefully chosen $\epsilon_a, \epsilon_b > 0$

7

Finding Confidence Interval



- Suppose X follows a distribution, given: $P(X \in [-1.96, 1.96]) = 0.95$
  - We are 95% sure that X will fall into the interval [-1.96, 1.96]
- If we find the distribution of $\hat{\mu} - \mu$, we can get the interval that has the probability as 95% (or 99%, can choose confidence level)
- Use $\hat{\mu}$ and the interval to calculate a range for μ, so that we are 95% sure μ fall into the range

Q: how to find the distribution of $\hat{\mu} - \mu$?

8

Confidence Intervals of the Normal Distribution

9

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2. Let $\hat{\mu} \coloneqq \frac{1}{n}\sum_i X_i$.

(Fact 1) $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

quiz candidate

$$\sqrt{n}\frac{\hat{\mu}-\mu}{\sigma} \sim N(0,1)$$

Recall:

- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under affine transformation (a and b constant):

$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$

(proof)

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Use this with $X = \sum_{i=1}^{n} X_i$, $a = \frac{1}{n}, b = 0.$

9

CDF of Normal Distribution

(Fact 2) If $Z \sim \mathcal{N}(0,1)$,

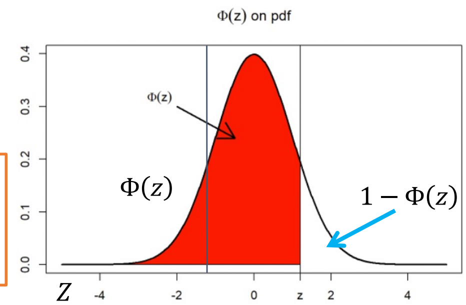
$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

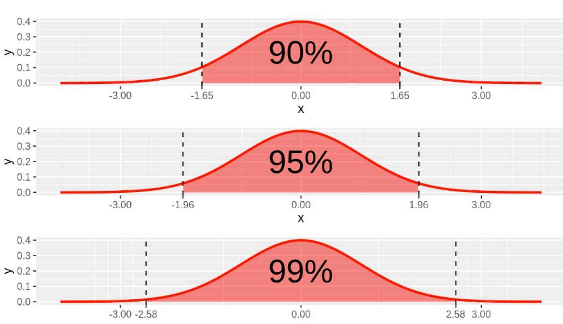
where $\Phi(z) \coloneqq \mathrm{P}(Z \leq z)$ is the CDF of Z.

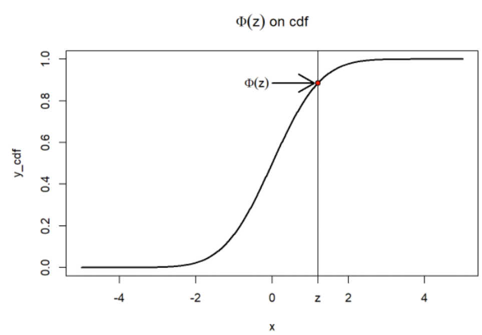
z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\sigma)^2}{2\sigma^2}} dx$$







10

Confidence Intervals of the Normal Distribution

Terminology: "standard" normal distribution := $\mathcal{N}(0,1)$

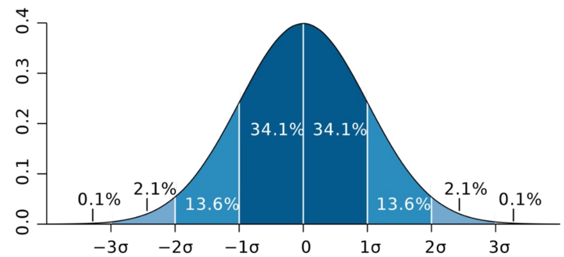
<u>(Fact 2)</u> If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) \coloneqq \mathrm{P}(\mathrm{Z} \le z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,



Gaussians almost do not have tails!

<u>(Corollary)</u>

$$P\left(\hat{\mu} \in \left[\mu - \frac{z\sigma}{\sqrt{n}}, \mu + \frac{z\sigma}{\sqrt{n}}\right]\right) = 1 - 2(1 - \Phi(z))$$

hints: use the fact $\sqrt{n}\frac{\hat{\mu}-\mu}{\sigma} \sim N(0,1)$. Set $Z \coloneqq \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$ and use Fact 2.

Confidence Intervals of the Normal Distribution

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2. Let $\hat{\mu} \coloneqq \frac{1}{n}\sum_i X_i$.

Fact 1 $$\sqrt{n}\,\frac{\widehat{\mu} - \mu}{\sigma} \sim N(0,1)$$

Fact 2 $$Z \sim N(0,1)$$
$$P(Z \in [-z, z]) = 1 - 2(1 - \phi(z))$$

$$Z \longrightarrow \sqrt{n}\,\frac{\widehat{\mu} - \mu}{\sigma}$$

z = 1.96: RHS ≈ .95, 95% confident
z = 2.58: RHS ≈ .99,

$$P\left(\sqrt{n}\,\frac{\widehat{\mu} - \mu}{\sigma} \in [-z, z]\right) = 1 - 2(1 - \phi(z))$$

$$P\left(\sqrt{n}\,\frac{\widehat{\mu} - \mu}{\sigma} \in [-z, z]\right) = P\left(\widehat{\mu} \in \left[\mu - \frac{z\sigma}{\sqrt{n}}, \mu + \frac{z\sigma}{\sqrt{n}}\right]\right)$$

Confidence Intervals of the Normal Distribution

13

note we can switch $\hat{\mu}$ and μ
$P\left(\mu \in \left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$

Finally, by our corollary,

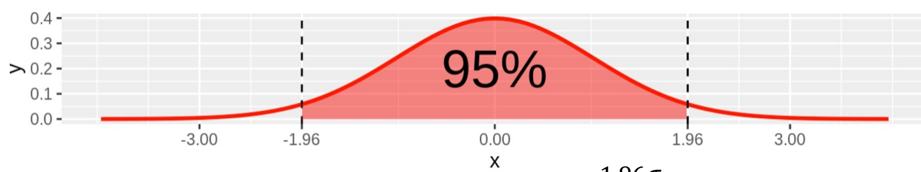
$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$$
$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \geq 0.99$$

$\hat{\mu} \in [\mu - 3, \mu + 3]$

$\mu - 3 \leqslant \hat{\mu} \leqslant \mu + 3$
$\hat{\mu} - 3 \leqslant \mu \leqslant \hat{\mu} + 3$

This is a confidence bound for the mean μ !!

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!



Confidence interval: (a,b), where $a = \hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}$, and $b = \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}$

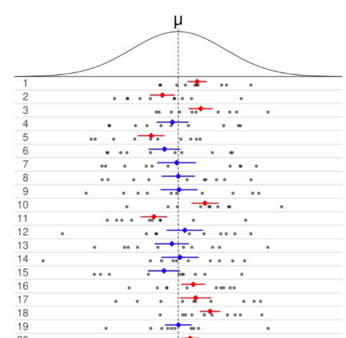
13

Caveat: interpreting confidence intervals

16

Recommended point of view:

- Assume: Heights of UA students follow a normal distribution $\mathcal{N}(\mu, 1)$ with unknown μ
- Fork m **parallel universes**. For each universe $u \in \{1,2,\dots,m\}$,
  - Subsample n UA students randomly, take the sample mean $\hat{\mu}^{(u)}$.
  - Compute the confidence bound $\left[\hat{\mu}^{(u)} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu}^{(u)} + \frac{1.96\sigma}{\sqrt{n}}\right]$



- The fraction of parallel universes where the random interval includes μ is *approximately* at least 0.95 if m is large enough.
- As m goes to infinity, the fraction will become arbitrarily close to a value that is at least 0.95.

16

Confidence bounds for arbitrary distributions

17

Recall: *If $X_1, \dots, X_n$ from an **arbitrary** distribution, can we still use the same method used for Gaussian?*

Short answer: YES, if n is large enough.

- Central limit theorem

$$\lim_{N\to\infty} \frac{\sqrt{N}}{\sigma}(\bar{X}_N - \mu) \to \mathcal{N}(0,1)$$

Q: What if n is not large enough (<30)?

17

Method 1: Gaussian (Corrected)

Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2.

(Fact 1) $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad \sqrt{n}\frac{\hat{\mu}-\mu}{\boxed{\sigma}} \sim \boxed{N(0,1)}$ → **Replaced by T-distribution**

$\widehat{\sigma}$ Sample STD

(Fact 2) If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) := \mathrm{P}(Z \le z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident
z = 2.58: RHS ≈ .99,

n≤30
small samples.

Let: $Z \longrightarrow \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \ge 0.95$$

$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \ge 0.99$$

Q: what if σ^2 is unknown and sample size is small (< 30)?

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

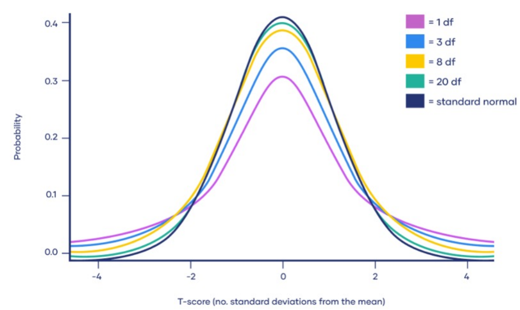
18

Method 1: Gaussian (Corrected)

Recall: Gaussian confidence interval with $\sqrt{n}\frac{\hat{\mu}_n-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
What if we use $\hat{\sigma}$ instead of σ?

(Theorem) $X_1, \dots, X_n$ with unknown μ, σ^2.
Let $\widehat{UVar}_n := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$ (unbiased version of sample variance). Then,

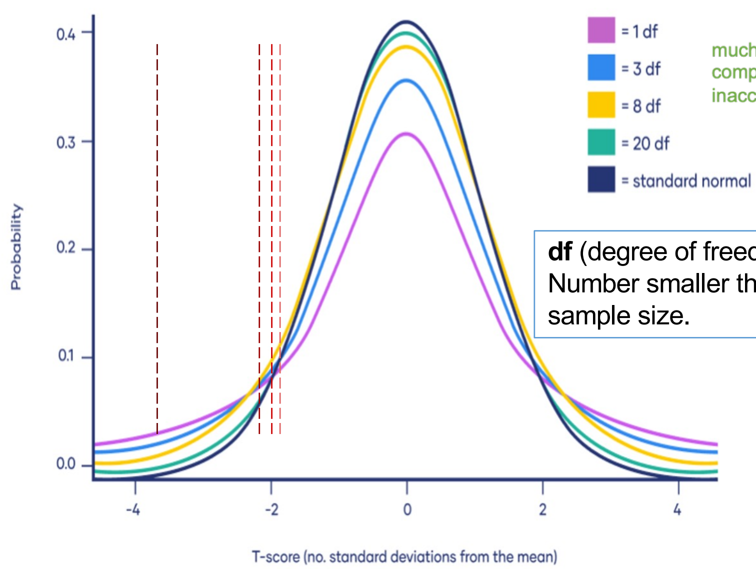
$$\sqrt{n}\frac{\hat{\mu}_n-\mu}{\sqrt{\widehat{UVar}_n}} \sim \text{student-t(mean 0, scale 1, degrees of freedom} = n-1)$$



σ

As df approaches infinity, T distribution becomes gaussian

T scores for different df



much larger number compensates for the inaccuracy of $\hat{\sigma}^2$

df (degree of freedom) = Number smaller than sample size.

(recall: 1.96 for gaussian)

```
import scipy.stats as st
alpha = 0.05
st.t.ppf(1-alpha/2,df=2)
=> 4.302652729911275

st.t.ppf(1-alpha/2,df=5)
=> 2.5705818366147395

st.t.ppf(1-alpha/2,df=10)
=> 2.2281388519649385

st.t.ppf(1-alpha/2,df=30)
=> 2.0422724563012373

st.t.ppf(1-alpha/2,df=100)
=> 1.9839715184496334
```

ppf=Percent Point Function

T Table

Degrees of freedom	Significance level 20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
......						
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∝	1.282	1.645	1.960	2.326	2.576	3.291

21

Method 1: Gaussian (Corrected)

22

With a similar derivation we have done before,
With at least 95% confidence:

$$\left[\hat{\mu} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Where $t_{\alpha/2,n-1}$ can be computed numerically.

Key take away: more conservative!
=> more likely to be correct.

***Common practice**: Apply this method even if we do not know whether true distribution is Gaussian.*

(recall: 1.96 for gaussian)

much larger number compensates for the inaccuracy of $\hat{\sigma}^2$

```
import scipy.stats as st
alpha = 0.05
st.t.ppf(1-alpha/2,df=2)
=> 4.302652729911275

st.t.ppf(1-alpha/2,df=5)
=> 2.5705818366147395

st.t.ppf(1-alpha/2,df=10)
=> 2.2281388519649385

st.t.ppf(1-alpha/2,df=30)
=> 2.0422724563012373

st.t.ppf(1-alpha/2,df=100)
=> 1.9839715184496334
```

Ppf=Percent Point Function

22

Method 2: Bootstrap

Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2.

<u>(Fact 1)</u> $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\sqrt{n}\frac{\boxed{\hat{\mu}-\mu}}{\sigma} \sim \mathcal{N}(0,1)$

Directly approximate distributions of $\hat{\mu} - \mu$

<u>(Fact 2)</u> If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) \coloneqq \mathrm{P}(\mathrm{Z} \leq z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,

Let: $Z \longrightarrow \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$$

$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \geq 0.99$$

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

23