Statistics: Estimation

1. Reminder: Continuous variables
2. Point estimation
3. Confidence intervals

Reminder: Continuous variables

Probability density function of x

$P(a \leq x \leq b) = \int_a^b f(x)\,\mathrm{d}x \quad \int_{min}^{max} f(x)\,\mathrm{d}x = 1$

Standard normal distribution $\mathcal{N}(0,1)$

- approx. 68% of values at $\mu \pm 1\sigma$
- approx. 95% of values at $\mu \pm 2\sigma$
- approx. 99% of values at $\mu \pm 3\sigma$

Standardized score

$Z = \frac{x-\mu}{\sigma}$

The main puzzle

	Notation	
Parameter	Sample	Population
Mean	$\bar{X}$	μ
Standard deviation	s	σ

The main solution: the properties of the standard normal distribution allow for statistical inference: the estimation, at a certain level of confidence, of the unobserved population parameters, using observed sample parameters.

Point estimation

Sample definitions

- the population mean μ is a population parameter
- the sample mean $\bar{X}$ is a point estimate of μ
- we know the sample n and its mean $\bar{X}$, but we do not know μ and might not know the true population N

Sampling error

- sampling variation causes $\bar{X}$ to vary
- the values of $\bar{X}$ form a sampling distribution
- its standard deviation $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean (SEM), which is estimated from the sample

CLT and LLN

Central Limit Theorem (CLT)

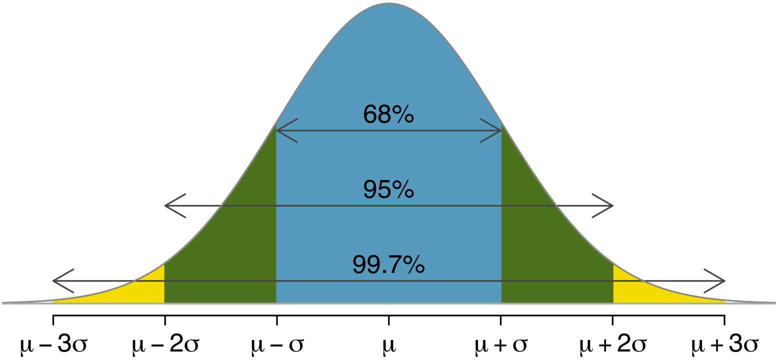
For ‘iid’ (independent and identically distributed) random variables $X_1, X_2, ..., X_n$, the sampling distribution of the mean approximates a normal distribution as $n > 30$ increases.

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\bar{X}_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2)$$

Law of Large Numbers (LLN)

$$\frac{X_1 + X_2 + ... + X_n}{n} = \mu$$

Standard normal distribution



Source: Diez *et al.* 2011

Standard normal distribution

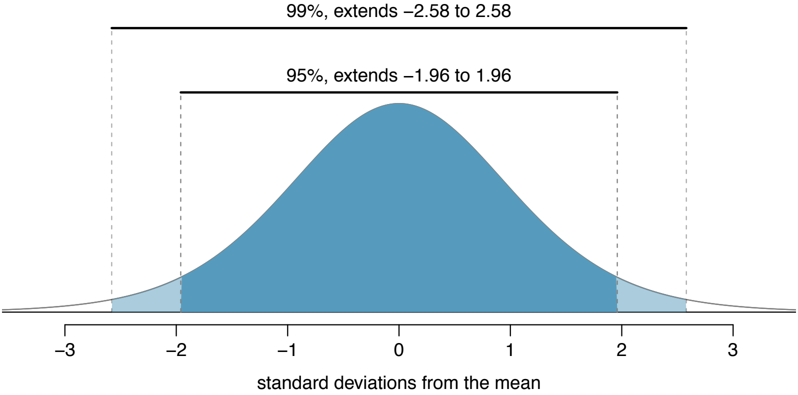


Figure 4.10: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

Confidence intervals

Confidence intervals

If the sampling distribution is approximately normal, fractions of the point estimates are contained within Z-scores:

- For a 95% CI: $\bar{X} - 1.96 \cdot SEM, \bar{X} + 1.96 \cdot SEM$
- For a 99% CI: $\bar{X} - 2.58 \cdot SEM, \bar{X} + 2.58 \cdot SEM$

Wider intervals trade precision for additional confidence.

Margin of error

The margin of error of the interval $\bar{X} \pm Z \cdot SEM$ is $Z \cdot SEM$.

Sanity check

Confidence intervals are estimations of the *population* parameter; they say nothing of the sample itself.

Homework

Read CK-12 handbook ch. 7–8 for next week

and enjoy the rest of your day.

Note: final stats exam will cover confidence intervals (Ch. 7) and hypothesis tests (Ch. 8). Histograms are part of the topic.