Class 15

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-02

Point Estimates and Population Parameters

Why can we use the ***sample statistic*** (e.g. sample mean $\bar{x}$, standard deviation s) as ***point estimates*** for the ***population parameters*** (e.g. population mean μ, population standard deviation σ)?

Building Models

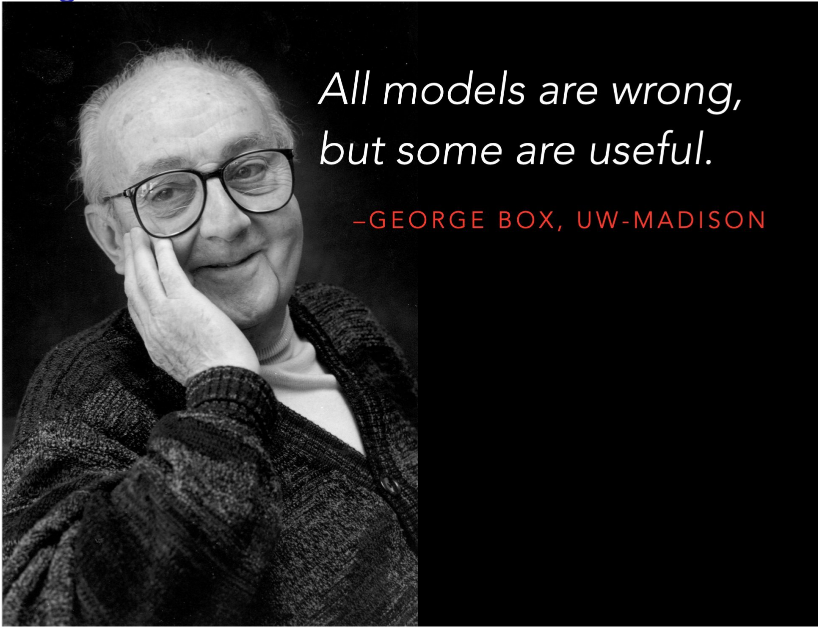


Figure 1: Statistical analysis requires making assumptions about the

What are you assuming? (The Model)

ASSUMPTION

The probability distribution of a random process follows a known distribution (e.g. a ***normal distribution***), which we can model and from which we can draw inferences about the parameters which govern that process.

What are you assuming? (Reliability)

ASSUMPTION

We have collected enough data and that data is trustworthy enough that our ***sample statistics*** are ***reliable*** estimators of the "ground truth" in our sample population.

What are you assuming? (Validity)

ASSUMPTION

Our sample population is sufficiently representative of our study population such that our ***sample statistics*** are ***valid*** estimators of the ***population parameters*** in our study population.

What are you assuming? (Generalizability)

ASSUMPTION

Our study population is sufficiently representative of our target population such that ***inferences*** about the ***population parameters*** of our study population are ***generalizable*** to our target population.

Accuracy vs Precision

- Accuracy describes how similar a sample statistic is to the "true" population parameter
- Precision describes how similar the sample statistics in a sampling distribution are to each other (i.e. the variability of the estimates)

Accuracy vs Precision

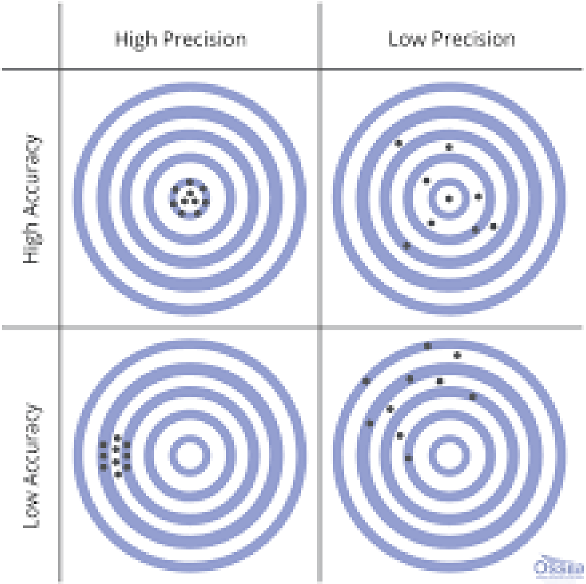


Figure 2: Contingency Table for Accurate and/or Precise Outcomes

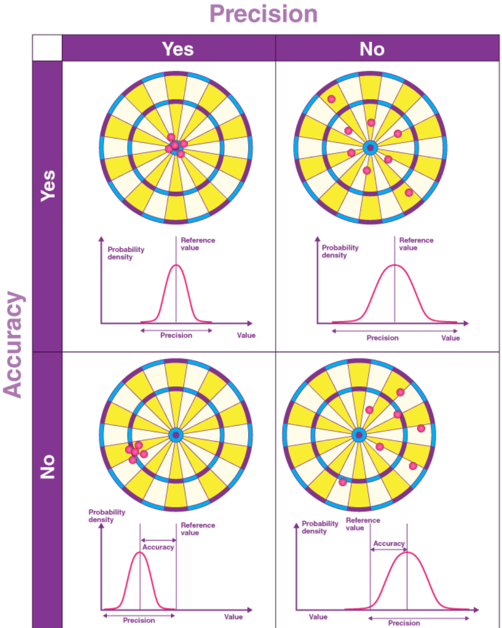
Why do we talk so much about study/sample/target populations?

- ***Reliable*** data $\rightarrow$ sample statistics are ***accurate*** estimators of sample population parameters
- ***Valid*** data $\rightarrow$ sample statistics are ***accurate*** estimators of sampling distribution in study population
- ***Generalizable*** data $\rightarrow$ sampling distribution of study population is ***accurate*** estimator of sampling distribution in target population

Why do we talk so much about study design?

- ***Larger samples*** $\rightarrow$ less variability $\rightarrow$ more ***precise*** estimates
- More representative samples $\rightarrow$ less biased estimates $\rightarrow$ more ***accurate*** estimates

Accuracy and Precision of Distributions



What is a confidence interval?

A ***confidence interval*** is a numerical range *inside* which a statistic is expected to occur with a given probability $1 - \alpha$ (alpha) in any theoretical sample from a given population

- $1 - \alpha$ is the ***confidence level*** and is often expressed as a %
- ***This is only true if your assumptions about the population hold.***

What is alpha (α)?

- α is called the ***confidence threshold***
- The statistic is expected to occur *outside* the ***confidence interval*** with probability α
- $(\alpha * 100)$% of confidence intervals for statistics from theoretical samples of this population will *NOT* contain the "true" population parameter
- A.K.A the ***Type I Error Rate*** or ***False Discovery Rate***

Point Estimates vs Confidence Intervals

- Point estimates are more ***precise*** than confidence intervals, but they are less likely to be ***accurate***
- Confidence intervals are more likely to be ***accurate*** than point estimates, but they are less ***precise***

Using both is best!

- A ***point estimate*** describes the ***location*** of an estimate or parameter's distribution
- A ***confidence interval*** describes the ***scale*** of an estimate or parameter's distribution
- The ***confidence threshold*** describes our uncertainty regarding these values

Choosing a Confidence Level

Choosing a ***confidence*** threshold α (alpha) is a trade-off between accuracy and precision.

- As confidence increases ($\alpha \to 0$), ***accuracy*** increases
- As confidence increases ($\alpha \to 0$), ***precision*** decreases

Example: Trade-Offs

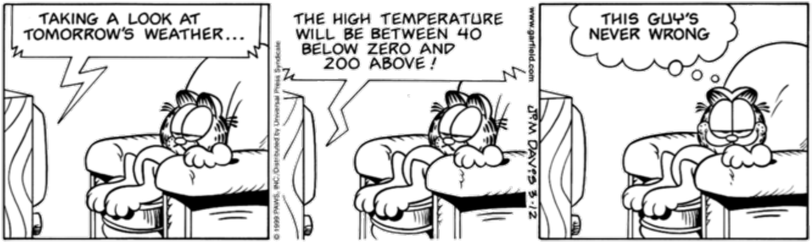


Figure 4: A weather forecast that is not very precise might accurately describe the weather on any given day, but it's certainly not very informative.

Practice: Confidence Intervals

Will a 95% confidence interval be wider (i.e. larger range) or narrower than a 90% confidence interval?

Practice: Confidence Intervals

Will a 95% confidence interval be wider (i.e. larger range) or narrower than a 90% confidence interval?

Wider

Practice: Precision

Which is a more ***precise*** estimator: a 95% or 90% confidence interval?

Practice: Precision

Which is a more ***precise*** estimator: a 95% or 90% confidence interval?

90% CI

Practice: Confidence Intervals

Will a 95% confidence interval be wider (i.e. larger range) or narrower than a 99% confidence interval?

Practice: Confidence Intervals

Will a 95% confidence interval be wider (i.e. larger range) or narrower than a 99% confidence interval?

Narrower

Practice: Accuracy

Which is more likely to be an ***accurate*** estimator: a 95% or 99% confidence interval?

Practice: Accuracy

Which is more likely to be an ***accurate*** estimator: a 95% or 99% confidence interval?

99% CI, ***if your assumptions hold***

How do we construct confidence intervals?

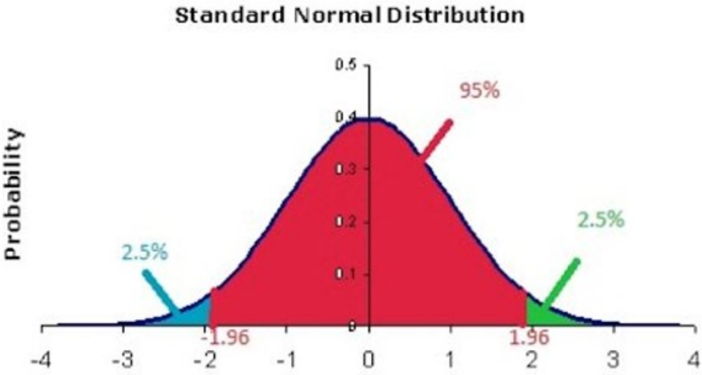


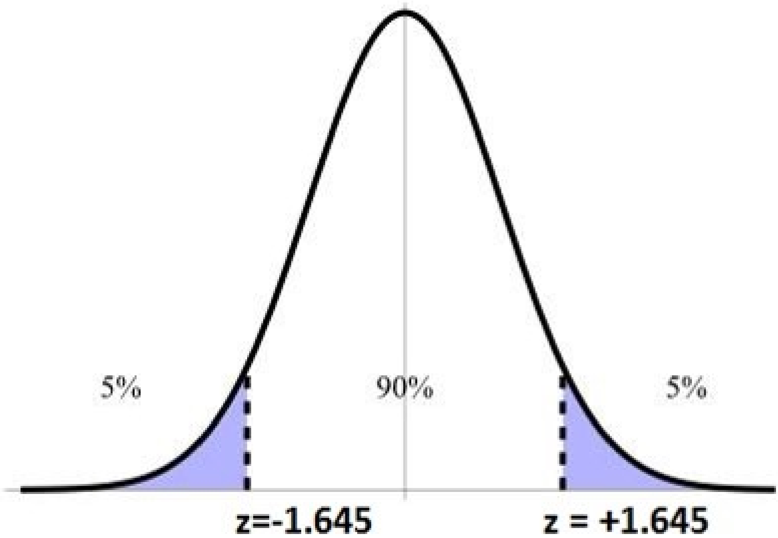
Figure 5: Properties of known distributions, like the 68-95-99.7 Rule, are used to calculate the bounds of a confidence interval.

Calculating a confidence interval

- A confidence interval is defined as $\text{point estimate} \pm \text{margin of error}$
- $\text{margin of error} = Z^* \times SE$
- $Z^* = \text{Z-Score}_{\alpha/2}$

Example: Z^* to $Z_{\alpha/2}$

If our confidence level is $1 - \alpha = 0.90$, then $\alpha = 0.1$.
$Z^* = \text{Z-Score}_{\alpha/2}$ and $\alpha/2 = 0.05$, so $Z^* = 1.645$.



Practice: Z^*

Which of the following Z-scores is the appropriate Z^* for constructing a 98% confidence interval?

1. $Z = 2.05$
2. $Z = 1.96$
3. $Z = 2.33$
4. $Z = 1.64$

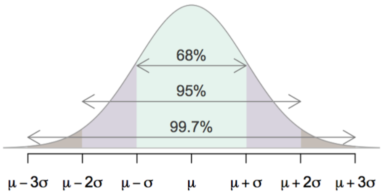
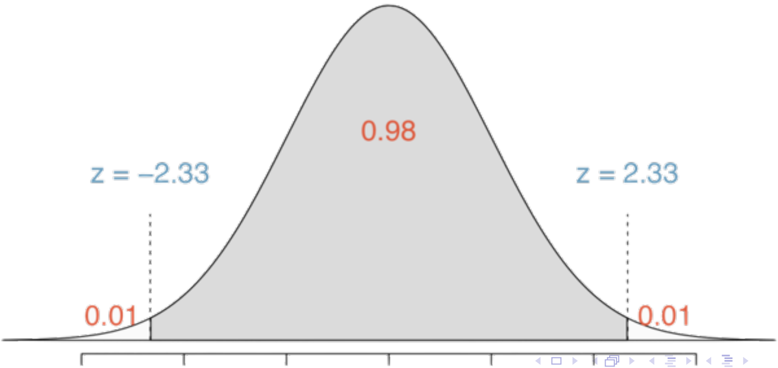


Figure 7: The 68-95-99.7 Rule for a Normal Distribution

Practice: Z^*

1. $Z = 2.05$
2. $Z = 1.96$
3. $Z = 2.33$
4. $Z = 1.64$



Example: Facebook Users

- Facebook is trying to assess the performance of their news feed algorithm based on whether or not users feel they are seeing relevant content.
- Their objective is to estimate the proportion of Facebook users who feel the algorithm works for them.
- They took a random sample of American Facebook users and asked if they think Facebook accurately categorizes their interests.
- 569 users out of the 850 sampled (67.5%) said they felt the algorithm was accurate.

Example: Populations

- What's the sample population?
- What's the study population?
- What's the target population?

We want to use ***reliable*** data from our sample to produce ***valid*** estimates of our study population distribution to make inferences that are ***generalizable*** to our target population.

Example: Calculating SE for a Proportion

$$
\begin{aligned}
SE_p &= \sqrt{\frac{p(1-p)}{n}} \\
&= \sqrt{\frac{0.67(1-0.67)}{850}} \\
&= 0.016
\end{aligned}
$$

Example: Calculating $Z^*_{0.95}$

For a 95% confidence interval, $1 - \alpha = 0.95$ and $\alpha = 0.05$, so $\alpha/2 = 0.025$. Therefore, $Z^*_{0.95} = Z_{0.025}$.

```
round(-qnorm(0.025, mean = 0, sd = 1), 2)
```

```
[1] 1.96
```

Our 95% confidence interval is defined by $0.67 \pm 1.96 \times 0.016$.

Example: Putting it Together

A 95% confidence interval for the proportion of all Facebook users who are satisfied with their algorithm is (0.64, 0.70).

```
Lower bound: 0.6383894
```

```
Upper bound: 0.7016106
```

Example: Interpretation

With 95% confidence, 64-70% of American Facebook users think Facebook categorizes their interests accurately...

Example: Interpretation

With 95% confidence, 64-70% of American Facebook users think Facebook categorizes their interests accurately...

Based on this study, with 95% confidence, we think the average percent of all Facebook users who are satisfied with their algorithm follows the distribution $N(0.67, 0.016)$...

Example: Interpretation

With 95% confidence, 64-70% of American Facebook users think Facebook categorizes their interests accurately...

Based on this study, with 95% confidence, we think the average percent of all Facebook users who are satisfied with their algorithm follows the distribution $N(0.67, 0.016)$...

...IF your assumptions are valid

What about confidence intervals for means?

Confidence intervals for means are calculated the same was as for proportions, but with the standard error of a mean calculation.

$$SE_{\mu} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$