Social Robotics

Albert-Ludwigs-Universität Freiburg

Felix Lindner, Laura Wächter, Bernhard Nebel

SoSe 2019

UNI FREIBURG

Inferential Statistics (Intro)

Current State

- You know how to look at your data.
- You know how to present your data.
- You got a first impression how to judge a data point as *extreme* or *usual* using IQR or z-Score.

The Problem

We face the problem that we want to investigate, whether some universally quantified statement holds, while we only have access to a subset of the overall population of entities the statement is quantifying over. This subset of the population we have access to is called the sample.

$\Rightarrow$Inferential statistics is about what we can reasonably say about the population given a sample.

Statistics vs. Parameters

	statistics	parameter
Mean	$\overline{X} = \frac{1}{N}\sum_i^N X_i$	$\mu = \frac{1}{N^*}\sum_i^{N^*} X_i$
Variance	$s^2_{biased} = \frac{1}{N}\sum_i^N (X_i - \overline{X})^2$ $s^2_{unbiased} = \frac{1}{N-1}\sum_i^N (X_i - \overline{X})^2$	$\sigma^2 = \frac{1}{N^*}\sum_i^{N^*} (X_i - \mu)^2$
Standard Deviation	$\sqrt{s^2}$	$\sqrt{\sigma^2}$

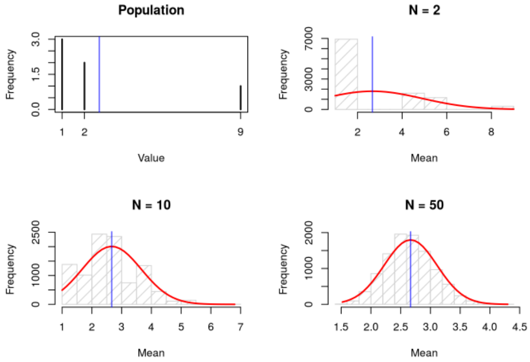
Sampling Distribution of the Sample Mean

UNI FREIBURG

The Gist

The sample mean will be approximately normally distributed for large sample sizes, regardless of the distribution from which we are sampling.

Evidence by Simulation



- Blue lines: Population mean μ.
- Grey Bars: Frequency of sampled means
- Red Gaussian: $\mathcal{N}(\mu, \frac{\sigma^2}{N})$

Mean of the Sampling Distribution of the Sample Mean

Let $X_1, \ldots, X_N$ be N independently drawn observations from a distribution with mean μ and variance σ^2. Thus, $E[X_i] = \mu$ for all i. Let's derive $E[\overline{X}]$, which we call the mean of the sampling distribution of the sample mean (also written as $\mu_{\overline{X}}$):

$$E[\overline{X}] = E[\frac{1}{N}\sum_i^N X_i] = \frac{1}{N}E[\sum_i^N X_i] = \frac{1}{N}\sum_i^N E[X_i] = \frac{1}{N}N\mu = \mu$$

Variance of the Sampling Distribution of the Sample Mean

UNI FREIBURG

Let $X_1, \ldots, X_N$ be N independently drawn observations from a distribution with mean μ and variance σ^2. Thus, $Var[X_i] = \sigma^2$ for all i. Let's derive $Var[\overline{X}]$, which we call the variance of the sampling distribution of the sample mean (also written as $\sigma^2_{\overline{X}}$):

$$Var[\overline{X}] = Var[\frac{1}{N}\sum_i^N X_i] = (\frac{1}{N})^2 Var[\sum_i^N X_i] = (\frac{1}{N})^2 \sum_i^N Var[x_i] = (\frac{1}{N})^2 N\sigma^2 = \frac{\sigma^2}{N}$$

- Hence, the standard deviation of the sampling distribution of the sample mean is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$.
- $\sigma_{\overline{X}}$ is also called the Standard Error.

Summary: Sampling Distribution of the Sample Mean

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

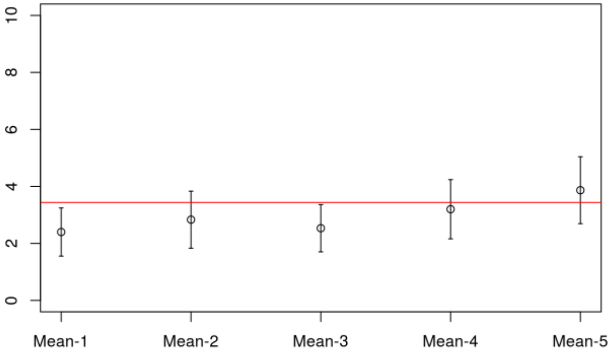
Application: Confidence Intervals I

- Suppose we know the population mean μ and standard deviation σ.
- Can we find boundaries within which we believe the mean of a sample of size N will fall with 95% probability?
- We know how our sample means are distributed, viz., $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$
  - The lower boundary $\overline{X}_{low}$ will be 1.96 standard errors below μ, and the upper boundary $\overline{X}_{up}$ will be 1.96 standard errors above μ.
  - $\mu - \overline{X}_{low} = 1.96 \times \frac{\sigma}{\sqrt{N}} \Rightarrow \overline{X}_{low} = \mu - 1.96 \times \frac{\sigma}{\sqrt{N}}$
  - $\overline{X}_{up} - \mu = 1.96 \times \frac{\sigma}{\sqrt{N}} \Rightarrow \overline{X}_{up} = \mu + 1.96 \times \frac{\sigma}{\sqrt{N}}$

Application: Confidence Intervals II

- Suppose we have collected some sample of size N, and we have computed the $\overline{X}$- and s^2-statistics.
- Can we find boundaries within which we believe the population mean μ will fall with 95% probability?
- We just look from the "sample's perspective".
- In need of parameters, we estimate $\mathcal{N}(\mu, \frac{\sigma^2}{N})$ by $\mathcal{N}(\overline{X}, \frac{s^2}{N})$ (which is okay, if $N > 30$).
  - The lower boundary X_{low} will be 1.96 standard errors below $\overline{X}$, and the upper boundary X_{up} will be 1.96 standard errors above $\overline{X}$.
  - $\overline{X} - X_{low} = 1.96 \times \frac{s}{\sqrt{N}} \Rightarrow X_{low} = \overline{X} - 1.96 \times \frac{s}{\sqrt{N}}$
  - $X_{up} - \overline{X} = 1.96 \times \frac{s}{\sqrt{N}} \Rightarrow X_{up} = \overline{X} + 1.96 \times \frac{s}{\sqrt{N}}$

Means and Confidence Intervals



- Red line: Population mean μ
- Dots: Sampled Means
- Lines through dots: 95% confidence intervals

Reporting Confidence Intervals

Report

We recorded the number of interactions with our robot per day for nine days (N = 9). The number of interactions ranged from 35 to 150 ($\overline{X}$ = 65.11, s = 33.59, 95% CI [43.16, 87.05]).

- Remember the data 35, 50, 50, 50, 56, 60, 60, 75, 150.

Application: Hypothesis Testing

- The sample mean has a distribution that is normal (for sufficiently large sample sizes), even when we are sampling from a distribution that is not normal.
- This is useful, because given μ and σ, we can compute the probability that some sample of size N with mean $\overline{X}$ stems from that population!
- We already know how we can judge whether some value from a normal distribution is ‘usual’ or rather ‘extreme’: z-Scores!
- Hence, we can judge a sample mean as ‘usual’ or ‘extreme’ by computing its z-Score.
- Let’s see how we can use this for hypothesis testing!

Very First Hypothesis Test

Suppose you have been deploying a robot (Robo-One) in your museum. You have recorded the number of interaction for a very long time, such that you can assume the collected mean and variance of the number of interactions to be the population mean $\mu_0 = 40$ and standard deviation $\sigma_0 = 4$. You have now bought a fancy new version of the robot, viz., Robo-Two. Your Hypothesis is that Robo-Two will generate much more interactions compared to Robo-One.

- Hypothesis H_1: Robo-Two generates more interactions than Robo-One.
- H_1 is of type (difference, directional)
- Can be written as $H_1 : \mu > \mu_0$, i.e., the population mean for interactions with Robo-Two (μ) is bigger than the population mean for interactions with Robo-One (μ_0), i.e., people generally interact more with Robo-Two than with Robo-One.

Very First Hypothesis Test

- The trick of inferential statistics is to first assume that the negation of H_1 is the case, which is called the Null-Hypothesis, written H_0.
- Then, we collect the data (viz., our sample)
- Subsequently, we show that our sample is so unlikely under H_0 that we are allowed to reject H_0 in favor of H_1.
  - In the example: $H_1 : \mu > \mu_0$, $H_0 : \mu \leq \mu_0$.

Very First Hypothesis Test

- Next, we record the number of interactions of Robo-Two for 16 days ($N = 16$), and we find a mean $\overline{X} = 42$.
- Given the population mean and standard deviation $\mu_0 = 40$ and $\sigma_0 = 4$, we know that the sampling distribution of the sample mean is $\mathcal{N}(40, \frac{16}{16})$.
- We compute the z-Score to assess how far our sample mean 42 is from the mean of the sampling distribution of the sample mean, 40: $z = (42 - 40)/\frac{4}{4} = (42 - 40) = 2$.

Very First Hypothesis Test

- Thus, observing a sampling mean of at least 42 under the assumption that the population mean is $\mu_0 = 40$ and the population standard deviation is $\sigma_0 = 4$ is as probable as $P(z \geq 2) = 1 - P(z < 2) = 0.0228$.
- Things will become even worse if we consider population means smaller than μ_0. Therefore, if we assume a significance level of $\alpha = 0.05$, we have reason to reject H_0 in favor of H_1.

Very First Hypothesis Test

Report

The number of interactions with Robo-Two is significantly higher than the number of interactions with Robo-One ($z = 2.0, p = 0.0228$).

- Because the hypothesis was directional, we checked if the z-Score of $\overline{X}$ was $z_{.95} = 1.65$ or higher. This is called a one-tailed test. The p-Value is just the probability $P(z \geq 2.0) = 0.0228$. This is below the significance level $\alpha = 0.05$.

Second Hypothesis Test

- This time, our H_1 hypothesis was that there is a difference between Robo-One and Robo-Two: $H_1 : \mu \neq \mu_0$.
- The null-hypothesis then is $H_0 : \mu = \mu_0$.
- We will reject H_0, if μ is too low or too high. Thus, we split our 5% significance level into two (2.5% at the lower end, and 2.5% at the higher end).
- We thus check if the z-Value is below $z_{.025} = -1.96$ or above $z_{.975} = 1.96$. This is a two-tailed test.
- As our z-Score was 2, we will also reject H_0 this time.

Very First Hypothesis Test

Report

The number of interactions with Robo-Two and with Robo-One differ significantly ($z = 2.0, p = 0.044$).

- Because the hypothesis was non-directional, we compute the probability to observe a z-Score at least as extreme as 2.0 (in both directions). The probability is thus $P(z \geq 2.0) + P(z \leq -2.0) = 0.0228 + 0.0228 = 0.0456$. This is below the significance level $\alpha = 0.05$.

Third Hypothesis Test

- This time, our H_1 hypothesis was that there there will be less interactions with Robo-Two than with Robo-One: $H_1 : \mu < \mu_0$.
- The null-hypothesis then is $H_0 : \mu \geq \mu_0$.
- We will reject H_0 if μ is too low. Thus, we test at the lower 5% tail, viz., if the z-Score is less or equal $z_{.05} = -1.65$.
- As our z-Score was 2, we will not reject H_0.

Very First Hypothesis Test

Report

The hypothesis H_1 stating that the number of interactions with Robo-Two will be less than with Robo-One was not supported ($z = 2.0, p = 0.9772$).

- This time we look only at the lower end, thus, we compute the probability $P(z \leq 2.0) = 0.9772$, which clearly is above the significance level $\alpha = 0.05$.

Type-I and Type-II Errors

- Our decisions to reject H_0 or not are based on probabilities! We see that our sample would be rather unusual if H_0 were true, thus we reject H_0. But it could be that we just had an unusual sample by chance. If we decide to reject H_0 although H_0 is actually true, then we commit a Type-I Error. Using the 5% significance level, we have a 5% chance per rejected H_0 hypothesis that we were wrong.
- If we instead reject H_1 although H_0 is wrong, then we commit a Type-II Error. This can happen, when there is an effect in the population, but our sample size was too small to detect that effect.

Final Note on z-Test

- Note that we have assumed that μ and σ^2 are known to us a-priori, or can be reasonably be approximated in case of a sufficiently big sample size.
- In many applications, we will not be able to enjoy this luxury.
- Therefore, we will learn about other test statistics, as well. But the main idea is the same, most of the time.

Sketches

Intentionally left blank :-)