Emory University
QTM 220 Regression Analysis
Learning Notes

Jiuru Lyu

January 31, 2024

Contents

1 Statistical Inference

1.1 Descriptive Statistics and Binary Covariates

Definition 1.1.1 (Location). The *location* of the data is where it is. It is about approximating the data by a constant.

$$Y_i \approx \mu, \quad \text{for } i = 1, \ldots, n$$

Example 1.1.2 D
ifferent ways to summarize location: mean, median

Definition 1.1.3 (Spread). The *spread* of the data is how far it tends to be from is location.
Definition 1.1.4 (Residuals). Spread summarizes the size of the *residuals* left over after constant approximation. We use $\hat{\varepsilon}$ to denote residuals.

$$\hat{\varepsilon}_i \coloneqq Y_i - \hat{\mu}.$$

Definition 1.1.5 (Median Absolute Deviation and Standard Deviation).

- The *median absolute deviation (MAD)* is the median size of residuals.
- The *standard deviation (sd)* is the square root of the mean squared size of residuals.

  Remark 1.1 *The standard deviation is a sort of average in which big residuals count more than smaller ones.*

Definition 1.1.6 (Distribution). We use *histograms* to summarize the *distribution* of the data.

Remark 1.2 *Distribution of the data tells us more information than location and spread, but less than dot plot.* For example, in this context, dot plot also include the identities of the individuals in addition to the number of people having salary in the range.

Definition 1.1.7 (Binary Data). *Binary data* only have two options, and we usually denote those two options as 1's and 0's.
Corollary 1.1.8 : Hence, when drawing a dot plot, everyone falls into either of the two lines representing 1 and 0.

Theorem 1.1.9 Location of Binary Data
The median is whichever outcome is the most common, and the mean is the proportion of 1's in the data.

Remark 1.3 *Hence, a histogram tells us no more information than $\hat{\mu}$.*

Theorem 1.1.10 Spread of Binary Data

- Median absolute deviation will always be 0 in a binary case.
- The standard deviation is the square root of the mean squared distance from the mean, and
$$\text{sd} = \sqrt{\hat{\mu}(1-\hat{\mu})}.$$

Proof 1. The claim concerning MAD is trivial. *Hint: there's only two possible values in the data, so median and MAD should always be the same.*

Now, let's consider the claim on standard deviation.

$$\begin{aligned}
\text{sd}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\mu})^2 \\
&= \frac{1}{n}\sum_{y:\{0,1\}}\sum_{i:Y_i=y}(Y_i-\hat{\mu})^2 \\
&= \frac{1}{n}\big\{N_1\big(1-\hat{\mu}^2\big)+(n-N_1)\big(0-\hat{\mu}^2\big)\big\} && [N_1 = \text{number of } 1\text{'s}] \\
&= \frac{1}{n}\big\{N_1\big(1-2\hat{\mu}+\hat{\mu}^2\big)+(n-N_1)\hat{\mu}^2\big\} \\
&= \frac{1}{n}\big\{N_1-2N_1\hat{\mu}+n\hat{\mu}^2\big\} \\
&= \frac{1}{n}\big\{n\hat{\mu}-2n\hat{\mu}\cdot\hat{\mu}+n\hat{\mu}^2\big\} && [N_1 = n\hat{\mu}] \\
&= \frac{1}{n}\big\{n\hat{\mu}-n\hat{\mu}^2\big\} \\
&= \hat{\mu}-\hat{\mu}^2 = \hat{\mu}(1-\hat{\mu}).
\end{aligned}$$

Therefore, we know
$$\text{sd} = \sqrt{\hat{\mu}(1-\hat{\mu})}.$$

■

Remark 1.4 *In binary data, knowing the mean $\equiv$ knowing everything else.*

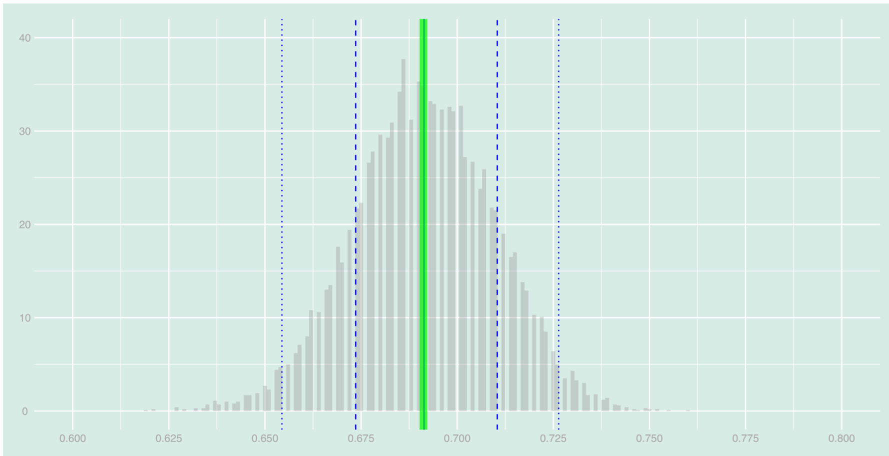
1.2 Population Inference for a Proportion

Definition 1.2.1 (Sampling Distribution). The *sampling distribution* is the distribution of estimates we'd get if we **replicated** our experiment over and over.

- Think of lots of people rolling the dice and reporting what they got.
- We consider this because it actually tells us something: it gives us an **interval** we can expect the proportion is in, and a statement about how much **confidence** we should have about it.

Example 1.2.2 Connecting Sample and Population

For each call i, we randomly select a voter with an id we'll call J_i. And we record as the call's outcome the turnout of the voter: $Y_i = y_{J_i}$. We can run this simulation using `R`.



- The *mean of the sampling distribution* is the solid blue line.
- The middle 2/3 of the sampling distribution lies between the dashed blue lines.
- The middle 95% of the sampling distribution lies between the dotted blue lines.
- Also, the population proportion is drawn as a wide green line.
- The question is "Could we predict how close we can get from the sampling before the election happened?" – Yes!
  - We will use an **interval estimate**: a *range of values* the population proportion is likely to be in.
  - The **width** of this interval speaks to the "how close" question.
  - The **coverage probability** (the probability we are right) qualifies this answer.
    * Our **point estimate** of the population proportion is the sample proportion $\overline{Y_n}$, where n is the size of the sample.
    * Now, we will try with some size of the interval. Say, x. Then, we are interested in the range of data $\overline{Y_n} \pm \frac{x}{2}$ (since the interval can be two-tailed).
    * Repeat the sampling process multiple times, say M times, and we notice that out of t times our interval "touches" the population proportion.
    * Then, we can define the coverage probability as follows:

$$\text{coverage probability} = \frac{t}{M} = \mathbf{P}\Big(\overline{Y}_n \in \overline{y}_N \pm \frac{x}{2}\Big),$$

where $\overline{Y}_n$ is our point estimate, $\overline{y}_N$ is the population proportion, and x is the width of the interval.

- Most of the time, we would like a 95% coverage probability, which means we will need to use a wider interval.
- Therefore, what we want to do is to choose a coverage probability and calculate the right width. An interval estimate like this (to ensure a given coverage) is called a **confidence interval**.
- The following figure shows a 95% coverage probability:

  - Our sample proportion 0.68 is close to the population proportion 0.69. Did we get luck? *No! In a million runs, almost all are within 0.05.*
  - Could we have predicted how close we would get before seeing the 0.69? *Yes! We can use a calibrated interval estimate – a Confidence Interval.*

- However, notice that this approach is not perfect: we cannot calibrate intervals like this in real life.
  - When we run our pool, we get a single point estimate $\overline{Y}_n$ based on our sample.
  - We don't know the sampling distribution of this point estimate until the election day.
  - However, what we actually do is almost the same: we will use an estimate of the sampling distribution in place of the thing itself.