Emory University

QTM 220 Regression Analysis

Learning Notes

Jiuru Lyu

February 28, 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 $\widehat{\varepsilon}$ to denote residuals.

$$\varepsilon_i \coloneqq Y_i - \widehat{\mu}.$$

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 1.8 : 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 $\widehat{\mu}$.*

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

$$\text{sd} = \sqrt{\widehat{\mu}(1-\widehat{\mu})}.$$

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.

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

Therefore, we know

$$\text{sd} = \sqrt{\widehat{\mu}(1-\widehat{\mu})}.$$

■

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

1.2 Population Inference for a Proportion

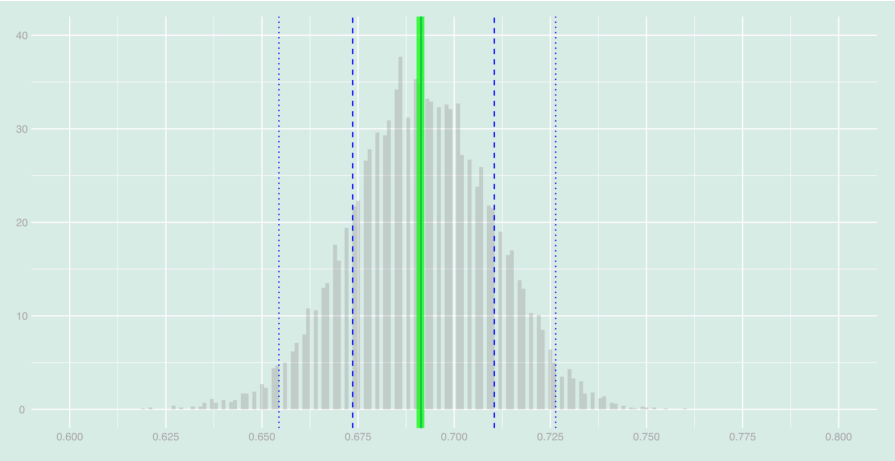
Definition 1.2.1 (Sampling Distribution). The *sampling distribution* is the distribution of estimates we'd get if we **replicated** our experiment over and over.

- 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.



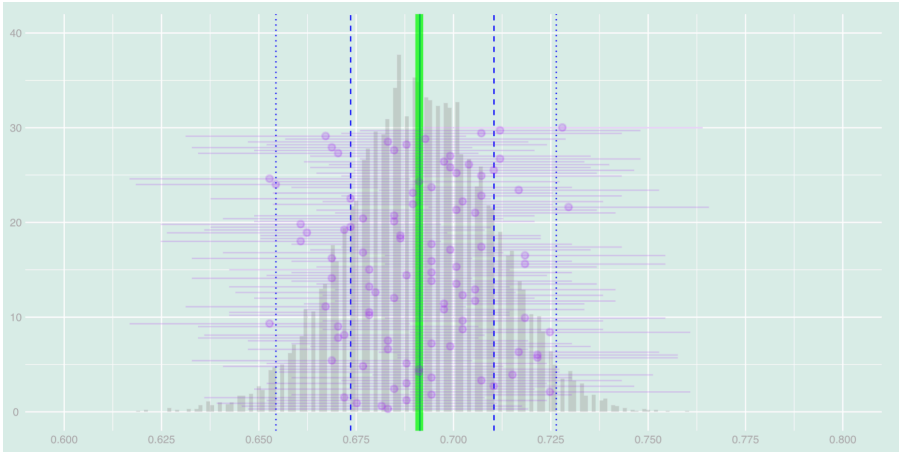
- The *mean of the sampling distribution* is the solid blue line.
- The middle 2/3 of the sampling distribution lies between the dashed blue lines.
- The middle 95% of the sampling distribution lies between the dotted blue lines.
- Also, the population proportion is drawn as a wide green line.
- The 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:

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

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.

1.3 Calibrating Interval Estimates

Theorem 1.3.1
An interval estimate covers the population proportion $\iff$ the corresponding point estimate is between the population proportion's arms.

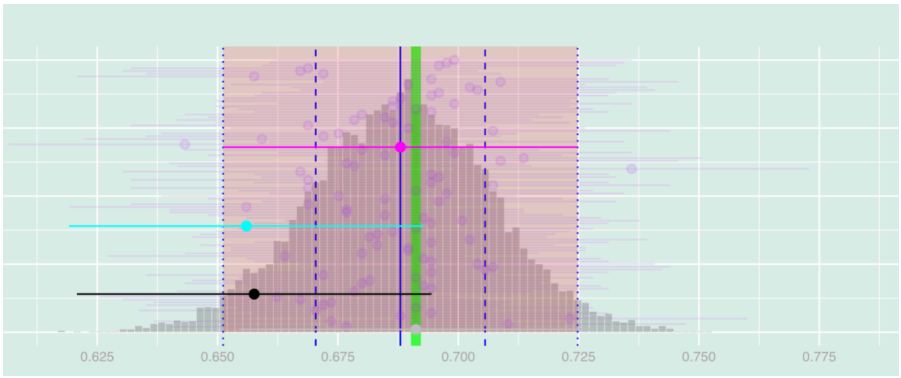
Remark 1.5 *With this Theorem, in Example 1.2.2, instead of looking at every interval and its arms to calculate coverage, we can draw arms of the same width around the population proportion.*

Equivalently, we can calculate the ***mass of the histogram*** *between the population proportion's arms.*

However, even with this Theorem, the problem still exists: unless we've seen the population, we cannot run simulations. Hence, in reality we will do calibration using an ***estimate of the sampling distribution****.*

Example 1.3.2

Here, we use our sample to estimate the sampling distribution. Compared with the actual population mean, the sample mean is a bit lower. Will this impact our estimation?



Solution 1.

It will not because we are not putting arms on draws from the estimated sampling distribution. We are putting arms on our point estimate, which is a draws from the actual sampling distribution. All that matters is the **width** of the estimated sampling distribution, and not the center. It turns out that the width calculated from the estimated sampling distribution and the population distribution should be the same (or close to the same). □

Theorem 1.3.3 Binomial Distribution Estimation
For a binary data type, we collected $Y_1, \cdots, Y_n$ as our sample. The distribution of the sample should follow a *Binomial Distribution*:

$$Y_i \sim \text{Binomial}(n, p).$$

Remark 1.6 *In theory, p should be calculated from the population. However, in the case of estimation, we will use our sample to estimate p. In the binary case, $p = \overline{Y}$.*

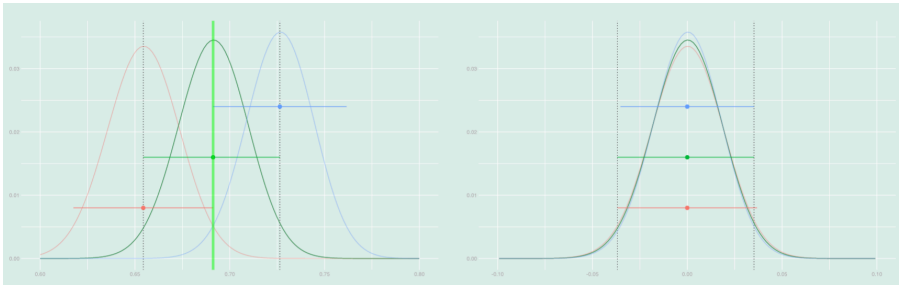
Binomial Distribution

```r
dbinom(x, n, p)
# x = number of heads, n = number of flips, p = probability of heads
```

Binomial Sample

```r
# To draw samples from estimated sampling distribution
samples <- rbinom(num, n, p)
# num = total number of draws,
# n = number of elements per drawing,
# p = probability of head
```

Example 1.3.4 How does the estimation work



The Binomial distribution is *continuous* as a function of p, so when p cahgnes little, the distribution changes little. That is to say that if we are not far off in the proportion, the estimated and actual sampling distributions are similar. The relevant difference (after centering) is even smaller because the way the binomial changes is mostly location.

This can be thought of a sort of "confidence interval" for our estimate of the sampling distribution. 95% of the time, we will get an estimate somewhere between the red and blue ones. As a result, a width of our interval estimate somewhere between the red and blue widths.

Example 1.3.5 The Bootstrap Interpretation

Let's revisit our distribution:

$$\text{Binomial}(n,p)/n.$$

This sample distribution indicates the proportion of 1's if we poll a sample of n object among whom the proportion of 1's is exactly $\overline{Y} = p$. This means that we can get a draw from our estimated sampling distribution by running a "poll" of the objects in our sample: rolling a n-sided die n times, calling up the corresponding object in our sample, and counting up the 1's we observe.

Boostraping Sample

```
bootstrap.samples = array(dim=10000)
for (rr in 1:10000) {
    Y.boot = Y[sample(1:n, n, replace=TRUE] # Boostraping
    bootstrap.samples[rr] = sum(Y.boot)/n
}
```

Remark 1.7 (Bootstrap Sampling) *The usual way of sampling is to use a sample to approximate the population. Since one sample can only generate one estimate, we need many samples. However, we cannot do this in reality. Instead, we only have one sample, so we will use bootstrapping. In that way, we resample the sample multiple times to general different estimates. Using the resampling distribution, we can eventually approximate the population.*

Definition 1.3.6 (Normal Distribution). The *normal distribution* is a function of two parameters: its mean and its standard deviation:

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Theorem 1.3.7 Central Limit Theorem

The sampling distribution and the bootstrap sampling distribution is approximately normal, and this is a consequence of the Central Limit Theorem.

1.4 Probability Review

Theorem 1.4.1 Properties of Expectations

Linearity of Expectation Suppose Y, Z are random variables and $a, b \in \mathbb{R}$. Then

$$\mathbf{E}(aY + bZ) = \mathbf{E}(aY) + \mathbf{E}(bZ) = a\mathbf{E}(Y) + b\mathbf{E}(Z).$$

Multiplication Rules Suppose Y, Z are independent ($\perp\!\!\!\perp$) random variables, then

$$\mathbf{E}(Y \cdot Z) = \mathbf{E}(Y) \cdot \mathbf{E}(Z).$$

Remark 1.8 *Without special notice, we assume random variables are independent in this course.*

Theorem 1.4.2 Variacne Decomposition
If $Y \perp\!\!\!\perp Z$, then $\mathbf{Var}(Y + Z) = \mathbf{Var}(Y) + \mathbf{Var}(Z)$.

Proof 1. Notice that

$$\begin{aligned}
\mathbf{Var}(Y+Z) &= \mathbf{E}[(Y+Z)^2] - \mathbf{E}(Y+Z)^2 \\
&= \mathbf{E}(Y^2 + Z^2 + 2YZ) - [\mathbf{E}(Y) + \mathbf{E}(Z)]^2 \\
&= \mathbf{E}(Y^2) + \mathbf{E}(Z^2) + 2\mathbf{E}(YZ) - \mathbf{E}(Y)^2 - \mathbf{E}(Z)^2 - 2\mathbf{E}(Y)\mathbf{E}(Z) && \text{[Linearity]} \\
&= (\mathbf{E}(Y^2) - \mathbf{E}(Y)^2) + (\mathbf{E}(Z^2) - \mathbf{E}(Z)^2) + 2\mathbf{E}(YZ) - 2\mathbf{E}(YZ) && \text{[Independence]} \\
&= \mathbf{Var}(Y) + \mathbf{Var}(Z).
\end{aligned}$$

■

Theorem 1.4.3 Binomial Expection and Variance
If $Y \sim \text{Binomial}(n, p)$, then

$$\mathbf{E}(\overline{Y}) = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i)\right] = p$$

and

$$\mathbf{Var}(\overline{Y}) = \frac{p(1-p)}{n}.$$

The proof of these two quantities are omitted.

Definition 1.4.4 (Conditional Expectation). The *conditional expectation* is $\mathbf{E}[Y \mid X = x]$, namely the expected value of Y given that $X = x$.

Theorem 1.4.5 Properties of Conditional Expectation

Law of Iterated Expectations for any random variables X and Y, we have

$$\mathbf{E}(Y) = \mathbf{E}\{\mathbf{E}(Y \mid X)\}.$$

Irrelevance of Independent Conditioning Variables When $Z \perp\!\!\!\perp X$ and Y, we have

$$\mathbf{E}(Y \mid X, Z) = \mathbf{E}(Y \mid X).$$

In other words, if Z is unrelated to X and Y, holding it constant does not affect the relationship between X and Y.