Statistical Analysis

A framework for producing knowledge from quantitative information

1. Formulate a question.
2. Design a study and collect data.
3. Choose a statistical model for the data.
4. Use the data to estimate and draw inferences about model parameters.
5. Make a judgment about the answer to the question.

These tasks are interrelated

Running example:

1. Question: Is a beer bottling company actually filling their bottles up to 12 ounces?
2. Study: take a sample of bottles from the line, measure the volume of beer from each bottle.

To discuss: What considerations do we need to make in this study?

Data: $y_1, y_2, \ldots, y_n$ are volumes of beer from n bottles.

What type of answer to our question can we hope to obtain from this study?

"It is plausible that macrobrewery ABC fills their bottles to 12oz. on average"

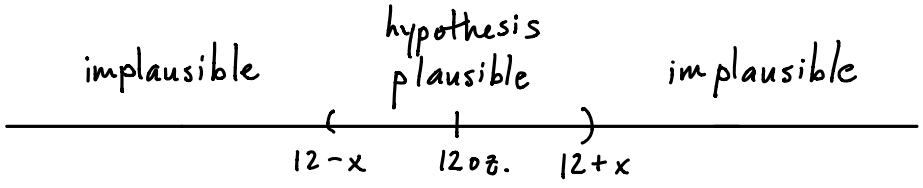
"It is not plausible that macrobrewery ABC fills their bottles to 12oz. on average."

A statement about the plausibility of a hypothesis given data is called a statisical inference.

Statistical hypothesis testing uses inductive logic and probability to turn this into a formal process.

High-level overview of how we make decisions about plausibility

1. before the experiment, pick a decision rule:
   "I will deem the statement implausible if my sample average is more than x ounces away from 12oz."



2. Use a statistical model to determine what x should be in order to guarantee we don't make the wrong decision too often.

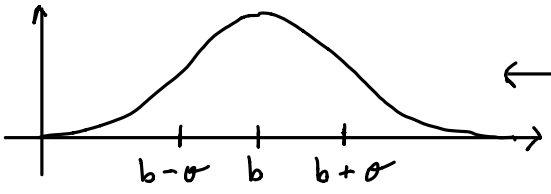
Statistical Models

Family of probability distributions meant to represent assumptions about how the data were generated.

We decide what we want to assume to be true, and what we want to learn from the data.

Bottling example: model the data y_i with random variable (RV) Y_i

$$Y_i \sim N(b, \sigma^2) \quad \leftarrow \text{Normal with mean } b, \text{ variance } \sigma^2$$



← probability density function.

RVs have probability distributions, are not specific numbers.

We can say $y_7 = 12.3$ but not $Y_7 = 12.3$

Notation: we try (but sometimes fail) to use lowercase letters for non-random quantities, uppercase for RVs.

To fully specify a probability distribution, we need to say how the random variables are related.

For now, assume that $Y_1, \ldots, Y_n$ are independent

- knowledge about Y_1 tells us nothing about $Y_2, \ldots, Y_n$

$$Y_1, \ldots, Y_n \overset{ind}{\sim} N(b, \sigma^2) \quad \text{or} \quad Y_i \overset{ind}{\sim} N(b, \sigma^2)$$

Sampling distributions

recall that our model for y_i is $Y_i \overset{ind}{\sim} N(b, \sigma^2)$

our goal is to say something about the parameter b.

We can estimate b using the data:

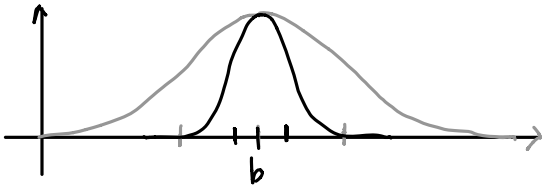
$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i$$ ← sample average of data

How close is $\hat{b}$ to b, the mean of the distribution?

The model helps us here. We can figure out the probability distribution of

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$ ← sample average of random variables

Turns out that $\hat{B} \sim N(b, \frac{\sigma^2}{n})$



This means that, under our model assumptions, $\hat{b}$ is unlikely to be more than $2\frac{\sigma}{\sqrt{n}}$ away from b.

Logical flip: b is unlikely to be more than $2\frac{\sigma}{\sqrt{n}}$ away from $\hat{b}$

Null hypothesis significance testing

data: $y_1, \ldots, y_n$

statistical model: $Y_i \sim P_\theta$ (some probability distribution with parameter θ)

Hypothesis: $H_0 =$ "$\theta = \theta_0$" (specific value: e.g. $\theta = 12$)

statistic: $u =$ function of $y_1, \ldots, y_n$

significance level: α, α between 0 and 1.

Decision Rule: reject or fail to reject hypothesis based on the value of the statistic.

Decision rule is constructed so that the probability of rejecting the hypothesis when it is true is equal to α. We use the sampling distribution of the statistic to calculate probabilities.

Bottling example:

data: $y_1, \ldots, y_{100}$. y_i = volume of beer in i^{th} bottle.

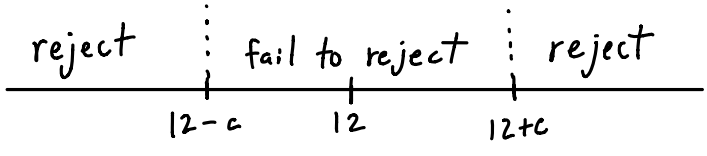
Model: $Y_i \overset{ind}{\sim} N(b, 0.1^2)$

Hypothesis: $H_0 =$ "$b = 12$"

statistic: $u = \frac{1}{n}\sum_{i=1}^{n} y_i$, a.k.a. $\bar{y}$

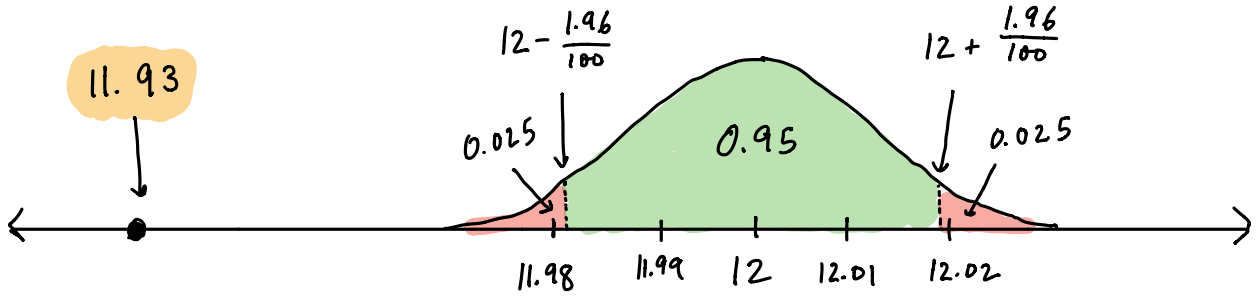
significance level: $\alpha = 0.05$

Decision rule: reject hypothesis if $|u - 12| > c$



We still need to figure out what c should be.

$$u = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad U = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N\left(b, \frac{0.1^2}{100}\right)$$



Result: reject the hypothesis that $b = 12$

We are saying that the hypothesis "$b = 12$" is implausible given the data we observed under the model assumptions we made

If we had observed $\bar{y} = 11.995$, the hypothesis would be plausible.

Confidence intervals

The confidence interval is the answer to the question:

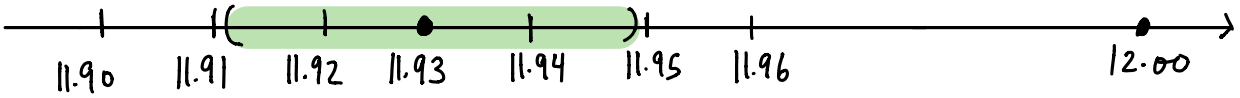
Which hypotheses are plausible?

It gives us a range of plausible values of the parameter.

Here's how it works:

Set up a hypothesis test for $\theta = \theta_0$

ci = { all values θ_0 for which we fail to reject $\theta = \theta_0$ }



$H_0: b = 12.00 \longrightarrow$ reject $\longrightarrow$ outside conf. interval

$H_0: b = 11.92 \longrightarrow$ fail to reject $\longrightarrow$ inside conf. interval

$H_0: b = 11.96 \longrightarrow$ reject $\longrightarrow$ outside conf. interval

$P(b_T \text{ inside C.I.})$

$= P(\text{fail to reject } b = b_T)$

$= 1 - P(\text{reject } b = b_T)$

$= 1 - \alpha$

b_T = "true value" of parameter

C.I. = random version of c.i.

This is why we call it a $100(1-\alpha)\%$ c.i. (e.g. 95% c.i.)

p-values:

Often our decision rule looks something like this:

reject H_0 if $u > c \longrightarrow$ more "extreme" than cutoff c

example: $u = |\bar{y} - 12.00|$. reject if $u > .0196$

The p-value is the probability, under the null hypothesis, of getting a statistic as extreme or more extreme than the one we actually observed.

$$\text{p-value} = P(U \geq u) \text{ under } H_0$$

U is random variable version of statistic u.

Bottling example:

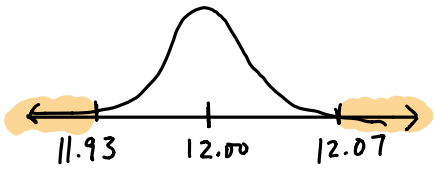
$u = |\bar{y} - 12.00| = |11.93 - 12.00| = 0.07$

decision rule: reject if $u > 0.0196$

$U = |\bar{Y} - 12.00|$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(12.00, 0.01^2)$

$P(U \geq u)$

$= P(U \geq 0.07)$



$= P(|\bar{Y} - 12.00| \geq 0.07)$

$= P(\bar{Y} \leq 11.93 \text{ or } \bar{Y} \geq 12.07)$

$= P(\bar{Y} \leq 11.93) + P(\bar{Y} \geq 0.07)$

$=$ pnorm(11.93, mean = 12, sd = 0.01, lower.tail = TRUE)

+ pnorm(12.07, mean = 12, sd = 0.01, lower.tail = FALSE)

$= 2.56 \times 10^{-12}$

Interpretation: the statistic we observed is very unlikely under the assumptions of the null hypothesis. Therefore we deem the null hypothesis to be implausible.

Connection to testing:

If the p-value is less than the significance level, we reject.

If the p-value is greater than the significance level, fail to reject.