Day Five Notes

Outline

1. Sampling Distributions
2. Binomial Setting and Sampling Distribution

Sampling Distributions

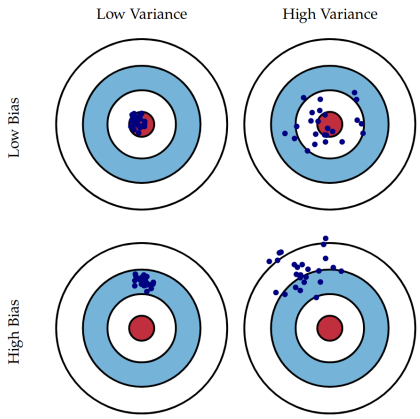

Fig. 1 Graphical illustration of bias and variance.

Figure 1: Illustration of bias and variance

Generally in science we have low bias and high variability.

Joke

A physicist, a biologist and a statistician go hunting:

They are hiding together in the bushes and they see a deer 70ft ahead of them. The physicist makes some calculations, aims and fires at the deer. His shot ends up 5ft to the left of the deer. The biologist analyzes the deer's movement, aims and fires. His shot ends up 5ft to the right of the deer. The statistician drops his rifle and happily shouts, "WE GOT IT!!"

In MATH-338, we assume our sample is generated using random sampling methods (simple random sampling)

- All samples of size `N` are equally likely

If we do not use good sampling methods:

- Probability distribution of sample changes $\rightarrow$ introduces bias

Sampling distribution depends on statistic & sample size

- For very large populations ($\sim$ 20x sample size)
- Sampling distribution does NOT depend on population size

Variability of sampling size $\downarrow$ as the sample size $\uparrow$.

Bias of sampling distribution $<<$ Bias introduced by bad sampling/study design

Variability of sampling distribution $<<$ Variability due to bad sampling/study design

Binomial (Probability) Distribution

Describe the number of "success" in `N` trials in the binomial setting

Four Conditions

Binary outcomes:

- All outcomes are classified either success or failure

Independent outcomes: (hand waved by good study design)

- The previous outcome does not influence the next outcome (flipping a coin, gender of a baby).

Number of outcomes = `N`

- Fixed sample size/number of trials
- Known in advance before any outcomes observed

Success is equally likely for each case/on each trial

- P = Probability of success = population probability of success
- the pass/fail rate of a class is 90/10, everyone is equally likely to fit these odds

Situations

- N term is violated : World Series games. A minimum four games is played but there is no fixed amount of games played
- I term is violated : teams who play away games vs home games

Binomial Random Variable

Let X = count(number) of success in a set of `N` outcomes obtained in the binomial setting

X has a PMF defined by:

$P(X = x) = \binom{n}{x} P^{\text{x}}(1 - P)^{\text{n-x}}$

- $n(x) = \frac{n!}{x!(n-x)!}$
  - Number of ways to get x success and n-x failures out of N trials
- probability of getting exactly x success
- Probability of getting n-x failures

Shorthand to: $X \sim B(n, p)$

> If you didn't understand them, an extreme simplification would be to say that you are repeating an activity with a chance of success p. Each repetition has the same chance of happening. This chance cannot be affected by the results of the other repetitions. You do this n times. X would represent the number of successes you got after doing n repetitions.
>
> For example, if you toss a fair coin 10 times, the number of heads you will get is:
>
> $X \sim B(10, 0.5)$
>
> because you toss it 10 times, and each toss has a 50% = 0.5 chance of being a head (because it is a fair coin).

Figure 2: Explanation

Example One

Toss a fair coin 8 times and let X = number of heads

Is X a binomial random variable?

- B: ✓success = heads
- I: ✓
- N: ✓n=8
- S: ✓p=0.5

$X \sim B(8, 0.5)$

Mean and Variance of a Binomial Random Variable

Consider the Bernoulli random variable:

```
X = {
    1: if outcome is success
    0: if outcome is failure
}
```

If p = probability of success, then:

$\mu_x = 1 * P + 0(1 - P) = P$

Binomial random variable is sum of `N` independent Bernoulli random variables

So if mean of binomial random variable = $P + P + P... + P = nP$ (n number of times)

Variance of Bernoulli random variable =

$(1 - P)^2 P + (0 - P)^2(1 - P) = P(1 - P)$

When X = 1 and X = 0 respectively ↑

$\Sigma(X - \mu)^2 P(X = x)$ $(1 - \mu)^2 P(X = 1)$

Variance of binomial random variable =

$P(1 - P) + P(1 - P) + ... + P(1 - P) = nP(1 - P)$ (n number of times)

Standard deviation if binomial random variable = $\sqrt{nP(1 - P)}$

Example Two (Question)

Sample of 2000 men get Gemfibrazil

Sample of 2000 men get some other drug (placebo)

It was assumed that 4% of men of the placebo group age get heart attacks (without drug intervention)

Looking at just placebo group:

- Define success, failure, `N`, `P` in this binomial setting
- Find mean, variance, and standard deviation of number of heart attacks the placebo group would experience.

Example Two (Answers)

Formulas used from section: Mean and Variance of a Binomial Random Variable

1

- Success is someone getting a heart attack
- Failure is someone not getting a heart attack
- 2000 is `N`
- 0.04 is `P`

2

- Mean: 80
- Variance: 76.8
- Standard deviation: 8.76356092

Distribution of Sample Proportion

$\hat{P} = \frac{X}{n}$

We have no idea how to estimate the number of successes in a very large population

Proportions are restricted to $[0, 1]$

Sample proportion and population proportion are on the same scale

$E[\hat{P}] = E[\frac{X}{n*X}] = \frac{X}{n} * E[X]$

If $X \sim B(n, P)$ then $E[X] = \mu_{\text{x}} = nP$

$E[\hat{P}] = \frac{X}{n}(n - P) = P$

$\hat{P}$ is an unbiased estimator of P

Variance$(\hat{P}) = V(\frac{1}{n} * X) = (\frac{1}{n})^2 V(x)$

If $X \sim B(n, P)$ then $V(x) = nP(1 - P)$

$V(\hat{P}) = (\frac{1}{n})^2(nP(1 - P)) = \frac{P(1-P)}{n}$

$SD(\hat{P}) = \sqrt{\frac{P(1-P)}{n}}$

Probability Problems involving p hat

$\hat{P}$ does NOT have a binomial distribution

However we can convert $X = n * \hat{P}$ and X has a binomial distribution