Contents

Day 18

Historical Motivation for T-Statistics

Example : Best Beer

Questions

- What type of barely is best?
  - How much barely can be produced?
  - How high quality is the barely?

Combine the two questions proposed into one variable → value per acre. We know there is variation in this variable. Different farms have variable soil, temperature, etc.

We need to know the average value per acre.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Problem is that we don't know μ nor σ

We can hand-waive this by saying:

- For sufficiently large samples we can replace population standard deviation σ with sample sample standard deviation s and our probability calculations are not too far off.

Our sample size is 4 farms, which is not enough.

For small samples, there is a lot variability in S. We need to take this variation into account.

Conclusions being Published

- Released in 1908 under the pseudonym Student who's real name is W.S. Gosset
- "Conclusions are not strictly applicable to populations known to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error."

Fisher was the only person who really cared about this report. This allowed Fisher for coming up with a better way. Most problems solved by statistics have been solved by beer and eugenics.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Definitions

Standard Error of the Mean (SEM)

The estimated standard deviation of the sampling distribution of $\bar{X}$, calculated as $\frac{s}{\sqrt{n}}$. More generally, the standard error of the statistic is the estimated (from sample data) standard deviation of its sampling distribution. As `n` ↑, SEM ↓

One-Sample t Statistic:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Has a (student's) t-distribution

With $n-1$ degrees of freedom, supposing a simple random sample of size `n` drawn from a population $\sim N(\mu, \sigma)$ where ($\mu\,known, \sigma\,unknown$). When population is not normal, procedures using t-distribution are still roughly accurate (unless clear skew/outlier issues). At high degrees of freedom, $t(n-1)$ is indistinguishable from $N(0,1)$ visually. It gets more rounded and less peaked at the top.

R Code

Will be added from the lab

```r
print("Hello World")
```

pt: values $\rightarrow$ probability qt: probability $\rightarrow$ values

Example

Suppose you have a simple random sample of 16, one-bedroom apartments for renting in a neighborhood. The sample mean rent price is $1280 per month and sample standard deviation is $180 per month.

Computational

- What is the SEM?
- What would the distribution of a t-statistic based on this sample?

Theoretical

- What else would you need to know/assume to calculate t-statistic?
- If you took a simple random sample of 25 apartments instead, would you express SEM to be larger or smaller? Why?
- Is it guaranteed?

Response

Computation

- $\frac{s}{\sqrt{n}} = \frac{180}{\sqrt{16}} = 45$
- Degrees of freedom: $n - 1 = 16 - 1 = 15 \rightarrow t(df = 15) \rightarrow t(15)$
- μ

Theoretical

- The SEM will go $\downarrow$ because there is less variabilit because there is less variability
- No because if $s > 225 \implies \frac{s}{\sqrt{n}} > 45$

One-Sample T-Statistic

$\mathrm{H}_0 : \mu = \mu_0$

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$\uparrow$ under the null hypothesis (H_0). Also, under H_0, $t \sim t(n-1)$

Neyman-Pearson Framework

Critical Region: on the t-scale

- $t \geq t$
- $t \leq -t$

$$t \geq \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Two-Sided Test

$$|t| \geq t$$

If $\mathrm{t}_{\text{observed}}$ is in critical region: accept H_1: $\mu = \mu_1$. Else accept H_0: $\mu = \mu_0$

$$t_{\text{observed}} = \frac{\bar{x_{\text{observed}}} - \mu_0}{\frac{s_0}{\sqrt{n}}}$$

Distributions we Should Know

- $N(\mu, \sigma)$
- $B(n, P)$
- $U(a, b)$ OR $V(a, b)$, can't tell
- $\chi^2(df)$
- $t(df)$