Robert Davidson
ST1112: Statistics

70% Exam
30% Continuous Assessment (3 parts)

Contents

1 Descriptive Statisitcs **3**
1.1 Sampling the mean . . . 3
Example 1.1 . . . 3
Definition 1.1 . . . 3
Definition 1.2 . . . 3
Definition 1.3 . . . 3

2 Interential Statistics - Interval Estimation **3**
2.1 Confidence Intervals . . . 3
2.1.1 Confidence Intervals for a mean . . . 3
Definition 2.1 . . . 3
2.1.2 Confidence Intervals for a small sample size . . . 4
Definition 2.2 . . . 4
Example 2.1 . . . 4
2.2 Bootstrap, proportions and counts . . . 5
2.2.1 Bootstrap . . . 5

3 Inferential Statistics - Hypothesis Testing **5**
3.1 Hypothesis Testing . . . 5
Definition 3.1 . . . 5
Definition 3.2 . . . 5
Definition 3.3 . . . 5
3.2 Core Definitions and Concepts . . . 6
Definition 3.4 . . . 6
Definition 3.5 . . . 6

1 Descriptive Statisitcs

1.1 Sampling the mean

In **probability** we consider the underlying process which has some randomness or uncertainity, and we try to figure out what happens

In **statistics** we consider the data that we have, and we try to figure out what the underlying process is. The basic aim to infer the population from the sample.

Example 1.1: Consider a jar of red and green jelly beans

A probabilist starts by knowing the proportion of red and green jelly beans in the jar, and then tries to figure out the probability of drawing a red jelly bean.
A statistician starts by drawing a sample of jelly beans from the jar, and then tries to figure out the proportion of red and green jelly beans in the jar.

Definition 1.1: Central Limit Theorem

Sample means follow a normal distrubution, centered on the popular mean, with a standard deviation equal to population standard deviation divided by the square root of the sample size.

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

Definition 1.2: Standard Error

The standard error is the variability in the sampling distrubution.
The standard error describes the typical difference between the sample measurement and the population parameter.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Definition 1.3: Estimate σ

Often the value of the population standard deviation is unknown, and hence the standard error of the mean is unknown.
We can estimate the value of the standard error using the sample standard deviation (s) as an unbiased estimator of the population standard deviation (σ).

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$$

2 Interential Statistics - Interval Estimation

2.1 Confidence Intervals

2.1.1 Confidence Intervals for a mean

Definition 2.1: Confidence Interval for $n > 30$

For a large sample size, $n > 30$, a Confidence Interval for the population mean is given by:

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

A 95% Confidence Interval, $\alpha = 0.05$, meaning we accept a 5% risk that our interval doesn't contain the true population mean.
This 5% is split into 2.5% in each tail of the distrubution : $Z_{\frac{\alpha}{2}} = Z_{0.025}$.
When using a normal table that shows "area to the left", we need to find the Z value that corresponds to

$1 - 0.025 = 0.975$. Thus: $Z_{0.025} = 1.96$.
95% Confidence is most commonly used because increasing the confidence level increases the width of the interval, this may not be useful.

2.1.2 Confidence Intervals for a small sample size

When σ is known, a 95% Confidence Interval for the population mean is given by:

$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

When σ is unknown, a 95% Confidence Interval for the population mean is given by:

$$\bar{X} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

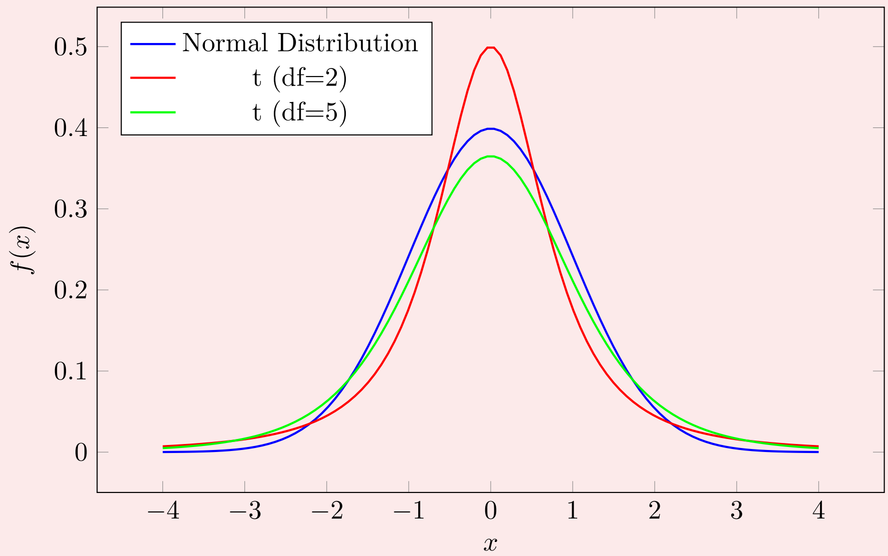
But, what if σ is unknown and $n < 30$?

Definition 2.2: Confidence Interval for $n < 30$

For a small sample size, $n < 30$, we use the **t-distribution instead of the normal distribution**. The t-distribution has heavier tails than the normal distribution and accounts for the additional uncertainty when estimating σ with s in small samples.
The confidence interval is given by:

$$\bar{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Where $t_{\frac{\alpha}{2}}$ is the critical value from the t-distribution with $n - 1$ **degrees of freedom**.
The degrees of freedom represent the *number of independent pieces of information used to estimate the standard deviation*. As the degrees of freedom increase, the t-distribution approaches the normal distribution - infinite degrees of freedom is the normal distribution.



Example 2.1: Find t-value from tables For a 95% Confidence Interval

We have:

$$a = 1 - 0.95 = 0.05 \Rightarrow \frac{a}{2} = 0.025$$

We also have:

$$n = 13 \Rightarrow df = 13 - 1$$

We want to find the $t_{12,0.025}$ value from the t-distribution table, which is 2.179.

2.2 Bootstrap, proportions and counts

What do we do when our data is not normally distributed? We have two options:

- Transform the data to make it normally distributed (e.g. log transformation, square root transformation)
- Use a non-parametric method (Bootstrap or CI for the population median)

2.2.1 Bootstrap

We can quantify the uncertainity in our estimate of the population mean by using the Central Limit Theorem or simulatiuon via the Bootstrap method, as follows:

1. Take a bootstrap sample - random sample taken with replacment from the original sample (same size as original sample)
2. Calculate the bootstrap statistic - such as mean, median, proportion, etc.
3. Repeat steps 1 and 2 many times
4. Calculate the bounds of the 95% Confidence interval as the middle 95% of the bootstrap distrubution

3 Inferential Statistics - Hypothesis Testing

3.1 Hypothesis Testing

Definition 3.1: Hypothesis Testing

A structured procedure to evaluate claims about a population paramater by comparing sample evidence against a proposed hypothesis

Definition 3.2: Null Hypothesis H_0

The defauly claim stating the population parameter equals a specific value (e.g. $\mu = 6.5$). It assumes that anu difference between the sample statistic and hypothesized value is due to random variation.

Definition 3.3: Alternative Hypothesis H_a

The competing claim the population paramater differs from the null hypothesis. It can be:

- **One-sided**: Claiming the paramater is either greater than or less than the specific value
- **Two-sided**: Claiming the paramater is not equal to the specific value

3.2 Core Definitions and Concepts

Definition 3.4: Test Statistic

A standardized measure that quantifies the diffence of the sample sample statisitic from the hypothesized population paramater, taking into account sample variability and size.

$$T_0 = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Where:

- $\overline{X}$ is the sample mean
- μ_0 is the hypothesized population mean
- s is the sample standard deviation
- n is the sample size

Definition 3.5: p-value

The probability of obsevering data as extreme (or more extreme than) the current sammple result if the null hypothesis is true. A smaller p-value indicates stronger evidence against H_0. It is not the probability that H_0 is true given the data.