OpenIntro Stats

Lydia Gibson

11/4/2021

Chapter 1: Introduction to data

Chapter 2: Summarizing Data

2.1 Examining numerical data

2.2 Considering categorical data

Chapter 3: Probability

3.4 Random Variables

3.4.1 Expectation

Random Variable

A random process or variable with a numerical outcome

Expected Value of a Discrete Random Variable

If X takes outcomes $x_1, ...x_k$ with probabilities $P(X = x_1), ..., P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by it's corresponding probability:

$E(X) = x_1 \times P(X = x_1) + ... + x_k + P(X = x_k)$

$= \sum_{i=1}^{k} x_i P(X = x_i)$

The Greek letter μ may be used in place of the notation E(X).

3.4.2 Variability in random variables

General Variance Formula

If X takes outcomes $x_1, ..., x_k$ with probabilities $P(X = x_1), ..., P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by Var(X) or the symbol σ^2, is

$\sigma^2 = (x_1 - \mu)^2 \times P(X = x_i) + ... + (x_k - \mu)^2 \times P(X = x_k)$

$= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j)$

The standard deviationof X, labeled σ, is the square root of the variance.

3.4.3 Linear combinations of random variables

Linear Combinations of Random Varibales and the Average Result

If X and Y are random variables, then the linear combination of the random varibales is given by

$aX = bY$

where a and b are fixed numbers. To compute the average value of a linear combination of random variables, plug in the average of each individula random variable and compute the result:

$a \times E(X) + b \times E(Y)$

Recall that the expected value is the same as the mean, e.g. $E(X) = \mu_X$

3.4.4 Variability in the linear combinations of random variables

Chapter 4: Distributions of random variables

4.1 Normal distribution

Chapter 5: Foundations for inference

5.1 Point estimates and sampling variability

5.2 Confidence Intervals for a proportion

5.3 Hypothesis testing for a proportion

Chapter 6: Inference for categorical data

6.1 Inference for a single proportion

6.1.1 Identifying when the sample proportion is nearly normal

Sampling Distribution of $\hat{p}$

The sampling distribution for $\hat{p}$ based on a sample of size n from a population with a true proportion p is nearly normal when:

1. The sample's observations are independent, e.g. are from a simple random sample.
2. We expect to see at least 10 successes and 10 failures in the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the success-failure condition.

When these conditions are met, then the sampling distribution of *hatp* is nearly normal with mean p and standard error $SE = \sqrt{\frac{p(1-p)}{n}}$.

6.1.2 Confidence Intervals for proportion

Confidence interval for a single proportion

$\hat{p} \pm z^* \times SE$

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the confidence interval:

Prepare: Identify $\hat{p}$ and n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one-proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure condition.

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Calculate. If the conditions hold, compute SE using $\hat{p}$ and z^*, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

6.1.3 Hypothesis testing for a proportion

Hypothesis testing for a single proportion

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list the hypothesis, identify the significance level and identify $\hat{p}$ and n.

Check. Verify conditions to ensure $\hat{p}$ is nearly normal under H_0. For one-proportion hypothesis tests, use the null value to check the success failure condition.

$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

Calculate. If the conditions hold, compute the standard error, again using p_0, compute the Z-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

6.1.4 When one or more conditions aren't met

For confidence intervals when the success-failure condition isn't met, we can use what's called the Clopper-Pearson interval.

6.1.5 Choosing a sample size when estimating a proportion

The *margin of error* is the part we add or subtract from the point estimate in a confidence interval.

The margin of error for a sample proportion is $z^* \sqrt{\frac{p(1-p)}{n}}$

$n =$

Difference of two proportions

6.2.1 Sampling distribution of the difference of two proportions

Conditions for the sampling distribution of $\hat{p_1} - \hat{p_2}$ to be normal

The difference $\hat{p_1} - \hat{p_2}$ can be modeled using a normal distribution when

- Independence, extended. The data are independent within and between the two groups. Genereally this is atisfied if the data come fromtwo independent random samples or if the dat come form a randomized experiment.
- Success-failure condition. The success-failure condition holds for both groups, where we check successes and failures in each group seperately.

When these conditions are satisfied, the standard error of $\hat{p_1} - \hat{p_2}$ is

$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

6.2.2 Confidence intervals for $p_1 - p_2$

We can apply the generic confidence interval formula for a differnce of two proportions, where we use $\hat{p}_1 - \hat{p}_2$ as the point estimate and substitute the SE formula:

$pointestimate \pm z^* \times SE \rightarrow \hat{p}_1 - \hat{p}_2 \pm z^* \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

6.2.3 Hypothesis tests for the difference of two proportions

Use the pooled proportion when H_0 is $p_1 - p_2 = 0$

When the null hypothesis is that the proportions are equal, use the pooled proportion ($\hat{p}_{pooled}$) to verify the success-failure condition and estimate the standard error:

$\hat{p}_{pooled} = \frac{numberof"sucesses"}{numberofcases} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$

Here $\hat{p}_1 n_1$ represents the number of succes in sample 1 since

$\hat{p}_1 = \frac{numberofsuccessesinsample1}{n_1}$

Similarly $\hat{p}_2 n_2$ represents the number of successes in sample 2.

6.2.4 More on 2-proportion hypothesis tests (special topic)

6.2.5 Examining the standard error formula (special topic)

6.3 Testing for goodness of fit using Chi-square

6.4 Testing for Independence in two-way tables

Chapter 7: Inference for Numerical data

7.1 One-sample means with the t-distribution

7.1.1 The sampling distribution of $\overline{x}$

Central limit Theorem for the Sample Mean

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\overline{x}$ will be nearly normal with

$Mean = \mu$ $StandardError(SE) = \frac{\sigma}{\sqrt{n}}$

7.1.2 Evaluating the two conditions required for modeling $\overline{x}$

*Independence.

*Normality.

Rules of Thumb: How to perform the normality check

There is no perfect way to check the normality condition, so instead we use two rules of thumb:

$n < 30$: If the sample n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

$n \geq 30$: If the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\overline{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

7.1.3 Introducing th t-distribution

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

Degrees of freedom (df)

The degrees of freedom describes the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When modeling $\overline{x}$ using the t-distrintuion, use df=n-1.

7.1.4 One sample t-confidence intervals

Finding a t-Conficence Interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$pointestimate \pm t^*_{df} \times SE \rightarrow \overline{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$

where $\overline{x}$ is the sample mean t^*_{df} corresponds to the confidence level and degrees of freedom df, and SE is the standard error as estimated by the sample.

Confidence interval for a single mean

Once you've determined a one-mean confidence interval would be helpful for an application, there are four steps to construction the interval:

Prepare. Identify $\overline{x}$, s, n, and determine what confidence level you wish ti use.

Check. Verify the conditions to ensure $\overline{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, find t^*_{df} , and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

7.1.5 One sample t-test

Hypothesis testing for a single mean

Once you've determined a one-mean hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list out the hypotheses, identify the significance level, and identify $\overline{x}$, s and n.

Check. Verify conditions to ensure $\overline{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, compute the T-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

7.2 Paired data

7.2.1 Paired observations

Paired Data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

7.2.2 Inference for Paired data

$H_0 : \mu_{diff} = 0$ $H_A : \mu_{diff} \neq 0$

$SE_{\overline{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}}$

$T = \frac{\overline{x}_{diff} - 0}{SE_{\overline{x}_{diff}}}$

7.3 Differnce of two means

7.3.1 Confidence interval for a differnce of means

Using the t-distribution for a differnce in means

The t-distribution can be used for inference when working with the standardized difference of two means if

Independence, extended. The data are indeprendent within and between the two groups, e.g. the data come from independent random samples or from a randomized experiment. Normality. We check the outlier rules of thumb for each group seperately.

The standard error may be computed as

$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

The official formula for the degrees of freedom is quite complex and is generally computed using software, so instead you may use the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom if software isn't readily available.

7.3.2 Hypothesis tests for the difference of the two means

7.3.3 Case study: two versions of a course exam

7.3.4 Pooled standard deviation estimate (special topic)

Pool standard deviations only after careful consideration

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.