BMS 617

Lecture 3 – Distributions, the confidence interval of a mean, and error bars

Distribution

• A distribution describes a data set.
  – Describes how frequently the values in the data set occur.
  – Can be used to describe an actual data set from a sample, or a theoretical data set in a population.
  • Can be used with discrete (categorical) or continuous variables.

Simple Distribution Example

• For a very simple example, we’ll use some categorical data from "Improving Adherence To A Mechanical Ventilation Weaning Protocol for Critically Ill Adults", American Journal of Critical Care, May 2006, 15;3.
  – Data gives number of patients in the study with each of a collection of health problems which lead to mechanical ventilation.
  – For a distribution, the number should be relative to the total
  – The total number of patients in this study was 129, so all numbers were divided by 129.

Comorbid Health Problem

Histogram of Ages

Distributions for continuous data

• What if the variable is continuous?
  – If measured precisely enough, no value occurs more than once...
  – Can graph the distribution with a histogram instead of a bar chart
  – Break the range of the variable into intervals and show the relative frequency in each interval
Population Distributions

- the distribution of the population, rather than the sample.
  - whole population.
  - However, we can sometimes deduce (mathematically) properties that must be true of its distribution.
  - In a lab experiment, the population is the set of all possible similar experiments we could perform.
  - consider it infinite.

Probability Density Functions

- A probability density function is like a histogram of relative frequency, but for an infinite number of data points!
  - It has the property that if we select any interval of values, the area under the graph in that interval is the probability (or relative frequency) a value will be in that interval.
  - The area under the whole graph must be equal to 1.
  - Theoretical distributions are usually described by their probability density functions.

Distributions for Experimental Data

- Think about any lab experiment which results in the measurement of a continuous variable.
- If we repeat the experiment many times under the same conditions, we will get similar, but not identical, results. What are the possible sources of the variation?
  - Imprecise measurement of reagents.
  - Imprecise pipetting
  - Nonhomogeneous mixes of solutions, suspensions, etc
  - For experiments based on animal or human samples, natural genetic variation, etc.
  - Many others...

Effect of Multiple Errors on Measurements

- measurement, they will often combine additively
- Most of the time, some variation will increase the measured value, and some will decrease it
  - So much of the variation cancels out, and most measured values will be close to the "true" value.
- Less often, all the sources of variation will combine to either increase or decrease the measurement
  - So some, but fewer, measured values will lie far from the true mean.
- In the 17th century, mathematicians determined what the distribution for experimental data would look like assuming there were many sources of variation and that they behaved additively.

The Gaussian Distribution

- The resulting distribution is called the Gaussian Distribution or Normal Distribution.

Mathematical Details of the Normal Distribution

- Remember, the probability a value lies in a given interval is the area under the curve over that interval.
- The formula for the distribution function is well-known.
- However, there is no way to get a formula for the area under an arbitrary portion of the curve.
  - Even with calculus!
- Instead, advanced numerical techniques are used, and values are published in tables or numerically calculated when required by software.
• If a variable is distributed according to the normal distribution, the chances that it lies within 1 standard deviation of the mean are 68.3%

• If a variable is distributed according to the normal distribution, the chances that it lies within 2 standard deviations of the mean are 95.5%

The probability values expressed in terms of standard deviations, as in the previous two slides, are true no matter what the standard deviation or mean are.

Consequently, if we know values for the normal distribution with mean 0 and standard deviation 1, we know the values for any normal distribution.

The normal distribution with mean \( \mu = 0 \) and standard deviation \( \sigma = 1 \) is called the standard normal distribution.

If a variable is normally distributed, then subtracting its mean and dividing by its standard deviation yields a variable with the standard normal distribution.

The normal distribution is particularly important in statistics for the following reason:

- Take some measurement, with any distribution whatsoever
- Sample that measurement times and compute the mean
- Repeat that many times to get lots of estimates of the mean
- Those means will be approximately normally distributed
- The approximation improves as \( n \) increases

Since many statistical tests are concerned only with the difference between means of values, tests that assume the normal distribution may work well even when the variable may not be normally distributed.

The argument that many sources of variation lead to a normal distribution relies on the assumption that the sources of variation are additive.

Sometimes the sources of variation are multiplicative

- In this case the effect is equally likely to double a value as to halve it.
- 100 is equally likely to be moved to 50 or 200
- Not symmetric!
- This leads to a lognormal distribution

The lognormal distribution
Calculating the CI of a mean

• Any statistical software worth using will perform this calculation for you. However, it's probably worth seeing how it works.
  - Compute the mean, \( \bar{m} \), and standard deviation \( s \) of your sample.
  - The confidence interval will be centered on the mean \( \bar{m} \), so we need to know how to compute the width of the interval.
  - The width of the interval, \( w \), which we call the margin of error, depends on a value from the T distribution, which we'll call \( t^* \).
  - \( t^* \) depends on:
    - \( n \): the sample size
    - \( \nu \): the degree of confidence
  - The margin of error, \( w = t^*s/\sqrt{n} \).

Example: CI of the mean

• From the mechanical ventilation paper referenced earlier:
  - Before the experimental intervention (which was to enhance adherence to clinical protocols), the mean duration of mechanical ventilation was 86.0 hours, with an SD of 68.0, and sample size 63.
  - After intervention, the mean was 70.8, sd 67.5, with a sample size of 66.
  - From tables, the \( t^* \) values for 95% confidence are 1.999 for 62 degrees of freedom and 1.997 for 65 degrees of freedom.
  - The margins of error are 17.1 (before intervention) and 16.6 (after)

Working with the lognormal distribution

• If the data are distributed lognormally:
  - The log of the data will be distributed normally
  - Take logs of all the data values and use standard statistical tests on the log data
  - Later in the course we'll see how to test whether data is distributed normally or not

Assumptions for computing the CI of a mean

• The calculation for the CI of a mean relies on four assumptions:
  1. Representative sample
  2. Independent observations
  3. Accurate data
  4. Population values are approximately normally distributed

The confidence interval of a mean

• number of samples, and compute the mean.
  - We interpret the mean of our samples as an approximation to the mean over the whole population.
  - We need to know the precision of this approximation
• Commonly, we will compute the mean from two or more groups of samples and compare them
  - Really interested in knowing the comparison of the means from the population
  - Cannot make this inference without a sense of the precision of our approximations
• Most intuitive way to do this is to compute the confidence interval for the mean.

Values determining the CI of a mean

• Four values are used in determining the CI of a mean:
  1. The mean of the sample.
  2. The SD of the sample.
  3. The sample size.
  4. The degree of confidence required.
• A higher degree of confidence requires a wider interval.

Populaton values are approximately normally distributed

Four assumptions:

1. Independent observations
2. Representative sample
3. Accurate data
4. Population values are approximately normally distributed

Accurate data

Independent observations

Representative sample

From the mechanical ventilation paper referenced earlier:

- Before the experimental intervention (which was to enhance adherence to clinical protocols), the mean duration of mechanical ventilation was 86.0 hours, with an SD of 68.0, and sample size 63.
- After intervention, the mean was 70.8, sd 67.5, with a sample size of 66.
- From tables, the * values for 95% confidence are 1.999 for 62 degrees of freedom and 1.997 for 65 degrees of freedom.
- The margins of error are 17.1 (before intervention) and 16.6 (after)
Example continued

- The confidence intervals for the before intervention mean thus has a lower limit of 86.0 - 17.1 = 68.9 and an upper limit of 86.0 + 17.1 = 103.1.
  - We say the 95% confidence interval for the mean is [68.9, 103.1].
  - Similarly, the 95% confidence interval for the mean duration of mechanical ventilation after intervention is [54.2, 87.4].
- Are you confident there is a real difference?

Confidence Intervals by Resampling

- The previous formula for computing confidence intervals assumes the population data is normally distributed.
- The resampling technique doesn’t make this assumption.
- Resample the sample data by randomly picking values from the sample. Do this so the resample is the same size as the sample.
  - Values from the sample may occur more than once in the resample.
  - This is called a pseudosample.
- Compute the mean of the resample.
- Repeat this many times (say, 1000), to get a distribution of means.
- Order the resampled means and identify the middle 95% (i.e. from the 2.5th percentile to the 97.5th percentile).
- This is an approximation to the 95% confidence interval for the mean.

Resampling

- Resampling can be used for many different data and statistics.
- No assumptions about the distribution
  - Requires only that the sample size is large enough to generate many (hundreds of) different resamples
  - Nine or ten values usually sufficient
- Reuses the same real sample to generate the pseudosamples - no real additional data is generated
- Theoretically valid approach
- But computationally intensive

The Central Limit Theorem (again)

- Recall the central limit theorem:
  - Suppose we have a population with mean $\mu$ and standard deviation $\sigma$ (but any distribution)
  - If we repeatedly take samples of size $n$ from this population, and compute their mean $m$:
    - The sample means $m$ will have a normal distribution (approximately)
    - The (population) mean of the sample means will also be $\mu$
    - The standard deviation of the sample means will be $\sigma/\sqrt{n}$.
  - Alternatively, we could say that the quantity $(\mu - m)/\sigma/\sqrt{n}$ follows the standard normal distribution.

Subtle but important point

- In the previous slide, the quantity that was stated to be normally distributed was
  \[
  \frac{\mu - m}{\sigma/\sqrt{n}}
  \]
- Notice the denominator includes $\sigma$, the population sd, not the sample sd.
  - Substituting the sample sd will not result in a normal distribution
  - Of course, in most cases we only know the sample sd, not the population sd.
Recall our standard setup:

- Suppose we have a normally distributed population, with mean μ.
- Repeatedly pick a sample of size n from this population.
- Compute the sample mean m, the sample SD s, and the quantity $t = (\mu - m)/(s/\sqrt{n})$.
- We want to know the distribution of values of t
- It turns out that the distribution depends on the sample size
- We characterize the t distribution by the degrees of freedom, which is n-1

For the t distribution with 65 degrees of freedom, the distribution depends on the sample size.

Mean $\mu$ is in the range $\mu \pm t^* s/\sqrt{n}$

For the t distribution with 65 degrees of freedom, the $t^*$ value is 1.997, as we saw earlier.

Interpreting the Standard Error of the Mean

- In a sense, it says: "If we computed lots of sample means from samples this big, this is our best estimate of how far they would be, on average, from the true population mean"
- The SEM does not quantify the amount of scatter within your sample
- If you have a large enough sample, the SEM will be very small, even if there is lots of scatter
- The SEM does, indirectly, quantify how precisely you know the population mean
- So does a confidence interval

Compute the sample mean m, the sample size n and
\[ t = (\mu - m)/(s/\sqrt{n}) \]
Why the SEM is hard to interpret

- When we work with normally distributed values, we can easily answer the question: "What proportion of values lie within one standard deviation of the mean?"
- The SEM is the standard deviation of sample means
- And sample means are not normally distributed because they are distributed according to the t-distribution, which itself depends on the sample size
- So the answer to "What proportion of sample means lie within one SEM of the true mean?" depends on the sample size
  - When n is very large, it is close to 68.3%
  - But for n=3, it is 57.7%

SEM and Confidence Intervals

- Both the SEM and Confidence Intervals quantify the precision with which we estimate the population mean.
  - Remember, the margin of error (half-width of the confidence interval) is \( w = t^* \times \frac{s}{\sqrt{n}} \)
    - which we can now write as \( w = t^* \times \text{(SEM)} \)
    - Here \( t^* \) is the value from the t-distribution so that the area from \(-t^*\) to \( t^* \) is the desired confidence level
    - If \( n \) is very large, then for a 95% confidence interval, \( t^* \approx 1.96 \)
    - But for \( n=4 \), \( t^* \approx 3.18 \), and for \( n=3 \), \( t^* \approx 4.30 \)

Error Bars

- When space is limited in an image (for example if you are showing multiple graphs in a figure), you may be restricted to showing just the sample mean and a single-value summary of the variation.
  - Again, if possible, show more detail
    - Column scatter plot, box and whisker plot, etc
- Common choices for the "error bars" are
  - Standard deviation
  - Standard Error of the Mean
  - Confidence Interval

Example: Gene Expression in Breast Cancer Cell Lines

- For an example, we'll use a recent experiment in which we measured expression of the gene GRHL2 in 51 different breast cancer cell lines
  - Cell lines categorized by basal type (Basal A, Basal B, Luminal)
  - Thirteen values for Basal A. Mean is 4.02, sd 1.58

Example Error Bars

Guidelines for using error bars

- Whatever you choose to plot, always clearly state what the error bars are
- Decide what you want to show:
  - To show variation among samples:
    - Use mean\(\pm\)sd or median and quartiles
  - To show precision of the estimation of the mean
    - Use mean and a confidence interval
    - Possibly use mean\(\pm\)SEM
      - but interpretation is far harder
Guidelines for reading error bars

• Make sure you know what the error bars represent
  – Should be stated in the figure legend
  – If not, might be buried in the methods somewhere…
• If the error bars show SEM:
  – Convert to SD with $s=\text{SEM}\sqrt{n}$
  – Or, if you are prepared to do more work, convert to a confidence interval with $w=t\times\text{SEM}$
  – Don’t assume $w=2\times\text{SEM}$ unless the sample size is in the hundreds