Intermediate Applied Statistics
STAT 460

Lecture 3, 9/8/2004

Instructor:
Aleksandra (Seša) Slavković
sna@stat.psu.edu

TA:
Wang Yu
wangyu@stat.psu.edu

---

Last Lecture

- EDA
  1. Measure of Center: Mean, Median, Mode
  2. Measure of spread (dispersion)
  3. Shape of distribution
     - Symmetry, skewness, modality, outliers

- Statistical displays
  - Histogram, density curve, boxplot, ...

- Valid, reliable, bias measurement

---

Review Example

- Price of Seagram 7 Crown Whisky (Chance 1991)
  - at 16 state-owned liquor stores:
    - 3.80 4.00 4.00 4.10 4.11 4.15 4.19 4.20
    - 4.20 4.50 4.55 4.65 4.74 4.75 5.05
  - at 26 privately owned liquor stores:
    - 4.29 4.29 4.50 4.54 4.55 4.75 4.75 4.75 4.79
    - 4.79 4.80 4.82 4.85 4.85 4.85 4.89 4.90
    - 4.95 4.95 4.95 5.10 5.20 5.25 5.29 5.30

- How would you describe this data?
What do we get if we draw a smooth curve over our histogram?

- Frequency curve=Density Curve
- This curve must satisfy: area under the curve equals 1

Frequency curve (contd.)

- Why do we care about them?
- They can tell us:
  - Proportion (or percent) of population for a variable that fall under a specific value
  - Proportion (or percent) will be above a specific value
  - Proportion (or percent) that will be between two certain values; in a certain range
  - The nth percentile = the position at which n% of the population will be below and (1-n)% will be above that position.

Who’s Gauss and why we might care about him?

- Normal curve=Bell-shaped Curve=Gaussian Curve
- Occurs so often in nature that it’s earned the name “normal”
- Any variable that is the sum or average of a small independent effects will have approximately a normal distribution (e.g. sample means)

Normal curves

- Symmetric, bell-shaped
- A normal curve is completely described by its mean and its standard deviation
- The mean determines the center of the distribution. Mean=Median=Mode.
- The standard deviation determines the shape of the curve. (It’s a distance from the mean to the change-of-curvature points on either side).
Normal curves

- What would happen if you change the mean of the normal distribution?
- What if you change the std.dev?

Proportions/Percentiles

- Suppose that when ordering shoes to restock the shelves in the store one knew that female shoe sizes were normally distributed with \( \mu = 7.0 \) and \( \sigma = 1.1 \). Don’t worry about where these values came from at this point, there will be plenty about that later. If the area under this distribution between 7.75 and 8.25 could be found, then one would know the proportion of size eight shoes to order. The values of 7.75 and 8.25 are the real limits of the interval of size eight shoes.

(ref. http://www.psychstat.msu.edu/introbook/ide121.html)

Standard Normal Distribution \( N(0,1) \)

- Mean = 0
- Variance = 1, hence St.dev = ?

Some notation:
- Let \( X \sim N(\mu, \sigma^2) \) (Variable \( X \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \))
- Let \( Z \sim N(0,1) \) where \( Z \) is known as **standardize score (Z-score)**
- Then \( Z = (X - \mu)/\sigma = (\text{observed value} - \text{mean})/(\text{st.dev}) \)

What we just did is known as transformation of variables
- If we know \( Z \) we can find \( X \) and vice versa (if we know value of \( X \) we can calculate the value of \( Z \))
- What is \( X \)?

Uses of \( Z \) ("Standardized") Scores

- Provide a measure of the relative standing of an observation independent of its original scale of measurement.
- If the mean GPA at Penn State is 2.9 with a standard deviation of 0.3, and you get a 3.5 then your \( z \) score is +2; you are 2 standard deviations above the mean. Assuming a normal distribution, this puts you above the 95th percentile (why?)
- An important advantage of converting to \( z \) scores is that the proportions under the standard normal distribution are found in tables in almost all statistics textbooks, and are very commonly used with many statistical procedures
Examples:

- What is the proportion of the Std. Normal population below $z = -0.41$?
- $z = 2.05$?
- What value of Std. Normal will 30% of the population fall below?
- What’s IQR of Std. Normal?
- What proportion of Std. Normal is between 2.05 and 0.15?
- What proportion is NOT between $z = 1.96 & z = 1.06$?
- What values of $z$ correspond to the middle area of 95%?

Examples:

- Now let $X \sim N(5, 10)$. What proportion is below 4?
- Suppose there is a normally distributed population whose standard deviation $\sigma$ is known to be (say) 10 but whose mean $\mu$ may not be known. How could we estimate $\mu$?
- Take a random sample of size $n$ = (say) 50.
- The sample mean $\bar{X}$ is a good estimator of the population mean $\mu$.
- According to the Law of Large Numbers, for a large enough sample the probability is very high that $\bar{X}$ is very close to $\mu$.
- We need some way to indicate an interval of plausible values – this will usually be the sample mean plus or minus a margin of error. How do we choose this interval?
The Empirical Rule

- The so-called “Empirical Rule” illustrates an important use of the standard deviation.

- If a population of measurements comes from a normal distribution with mean \( \mu \) and standard deviation \( \sigma \), then:
  - about 68\% of the scores are between \( \mu - \sigma \) and \( \mu + \sigma \)
  - slightly more than 95\% of the scores are between \( \mu - 2\sigma \) and \( \mu + 2\sigma \)
  - about 99.7\% of the scores are between \( \mu - 3\sigma \) and \( \mu + 3\sigma \)

Central Limit Theorem

- Distribution of a sample mean
  - If numerous samples of size \( n \) are taken, the frequency curve of the sample means (\( \bar{x} \)'s) from those various samples is approximately bell shaped with mean \( \mu \) and standard deviation \( \frac{\sigma}{\sqrt{n}} \).

- \( \bar{x} \sim N(\mu, \frac{\sigma^2}{n}) \)

- Holds if:
  - \( X \) is normally distributed
  - \( X \) is NOT normal, but \( n \) is large (e.g. \( n > 30 \)) and \( \mu \) finite,
  - No matter what the distribution of \( X \) is the above should hold even if we don’t know the distribution of \( X \).
  - For continuous and categorical (but proportions; later!!!)

Empirical Rule and CLT

- **Empirical Rule:** There is an approximately 95\% probability that a randomly selected observation from this distribution is in the interval \( (\mu - 2\sigma, \mu + 2\sigma) \).

- **Extension of Empirical Rule using Central Limit Theorem:** There is an approximately 95\% probability that the sample mean of a random sample of size \( n \) from this distribution is in the interval \( (\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}) \).

A Confidence Interval

- We are 95\% confident that \( \mu \) is in the interval \( (\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}) \)

- “95\% confident” does not mean that there is a 95\% probability that the true mean is in this interval.

- What does it mean?
Example

- Suppose we know that $\sigma = 10$, and we take a random sample of size 50 and get $\bar{x} = 8.3$.
- We can say with "95% confidence" that $\mu$ is somewhere in the interval.
- What's the distribution of the sample mean?
- What's the 95% CI? $\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$
  i.e. $8.3 \pm 2\frac{10}{\sqrt{50}}$
  i.e. $(5.47, 11.13)$

Generalizing the 95% Standard

Assuming that $\sigma$ is known, the multiplier for a $(1-\alpha)^{100\%}$ confidence interval is the $(1-\frac{\alpha}{2})^{100\%}$ percentile of the normal distribution.

Simplified Expression for a 95% Confidence Interval

There is a constant multiplier, usually a constant around 2 or a little higher, that comes from the distributions being used and the degree of confidence required.

The "margin of error" is some multiplier times the standard error, and it is added to and subtracted from the mean to get the endpoints of the interval.

The standard error of the mean, which is the standard deviation of the sampling distribution, is $\sigma/\sqrt{n}$.

Need for a Further Generalization

- What if we can’t assume that $\sigma$ is known?
- We use $s$ (the sample standard deviation) to estimate $\sigma$.
- If the sample is very large, we can treat $\sigma$ as known by assuming that $\sigma = s$. According to the law of large numbers, this is not too bad a thing to do.
- But if the sample is small, the fact that we have to estimate both the standard deviation and the mean adds extra uncertainty to our inference. In practice this means that we need a larger multiplier for the standard error.
- Next: t-distribution and t-test, hypothesis testing, statistical significance.
Applets

- [link](http://www.kuleuven.ac.be/ucs/java/index.htm)
- [link](http://bcw.whfreeman.com/ps4e/pages/bcs-main.asp?v=category&s=00010&n=99000&i=99010.01&n=)

_________________________________________