MATH 3339
Statistics for the Sciences
Live Lecture Help

James West
jdwest@uh.edu

University of Houston

Session 9

Office Hours: see schedule in the "Office Hours" channel on Teams
Course webpage: www.casa.uh.edu

Email policy

When you email me you **MUST** include the following

- MATH 3339 Section 20024 and a description of your issue in the **Subject Line**
- Your name and ID# in the **Body**
- Complete sentences, punctuation, and paragraph breaks
- Email messages to the class will be sent to your Exchange account (user@cougarnet.uh.edu)

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/

Outline

Updates and Announcements

- Test 1 total is "Test 1" + "Test 1 FR".
- Test 2 is in 4 weeks.
- I will replace one lower test grade with Final Exam grade.

PDF of a Normal Distribution

A continuous random variable X is said to have a **Normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

For all $-\infty < x < \infty$.

The cdf of X when $X \sim N(\mu, \sigma)$ is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/2\sigma^2} dt$$

Standard Normal Distribution

When $X \sim N(\mu, \sigma)$, we can standardize the values by forming:

$$Z = \frac{X - \mu}{\sigma}$$

where $\mu_Z = 0$ and $\sigma_Z = 1$ to get the pdf:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

The cdf of $Z \sim N(0, 1)$ is

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Normal Approximation to Binomial

Let X be a binomial random variable based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has an approximate Normal distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. In particular, for x = a possible value of X,

$$\begin{aligned} P(X \leq x) &= Binom(x; n, p) \\ &\approx \text{(area under the normal curve to the left of } x + 0.5) \\ &= \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

In practice, the approximation is adequate provided that both $np \geq 10$ and $n(1-p) \geq 10$.

Shape of the Sample Mean Distribution

- If a population has a Normal distribution, then the sample mean $\bar{X}$ of n independent observations also has a Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
- **Central limit theorem**: For *any* population, when n is large $(n \geq 30)$, the sampling distribution of the sample mean $\bar{X}$ is approximately a Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Notes about finding probabilities for $\bar{X}$

- We have a sample size n. Thus the standard deviation changes by that value $\mathrm{SD}(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
- The mean stays the same. $\mathrm{mean}(\bar{X}) = \mu_{\bar{X}} = \mu$.
- If we know that the original distribution is Normal **or** we have a large enough sample ($n \geq 30$). We can use the Normal distributions to find the probabilities.

Sample Proportions

- The population proportion is p a parameter. In some cases we do not know the population proportion, thus we use the sample proportion, $\hat{p}$ to estimate p.

- The sample proportion is calculated by: $\hat{p} = \frac{X}{n}$

- X = the number of observations of interest in the sample or the number of "successes" in the sample.

- n = the sample size or number of observations.

- Recall that $X \sim \text{Bin}(n, p)$ and can be approximated by the Normal distribution with $\mu_x = E(X) = np$ and $\sigma_X = SD(X) = \sqrt{np(1-p)}$ as long as $np \geq 10$ and $n(1-p) \geq 10$.

- Now we want to know how is $\hat{p} = \frac{X}{n}$ distributed. Thus we want to know $\mu_{\hat{p}} = E(\hat{p})$ and $\sigma_{\hat{p}} = SD(\hat{p})$.

Shape of the distribution of $\hat{p}$

We can use the **Normal distribution** as long as

- The sampled values must be random and independent of each other. This can be tested by **10% Condition**: The sample size must be no larger than 10% of the population.
- The sample size, n must be large enough. This can be be tested by **Success / Failure Condition**: The sample size has to be big enough so that both np and $n(1-p)$ at least 10.

Center of the distribution of $\hat{p}$

- The center is the mean (expected value): $\mu_{\hat{p}} = p$ the proportion of success.
- $\hat{p} = \frac{X}{n}$ where X is the number of **successes** out of n observations. Thus X has a binomial distribution with parameters n and p.
- The mean of X is:

$$\mu_X = E(X) = np$$

- Thus the mean of $\hat{p}$ is:

$$\mu_{\hat{p}} = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{\mu_X}{n} = \frac{np}{n} = p$$

Spread of the distribution of $\hat{p}$

- The spread is the standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.
- The variance of X is:

$$\sigma_X^2 = Var(X) = np(1-p)$$

- The variance of $\hat{p}$ is:

$$\sigma_{\hat{p}}^2 = Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Unbiased Estimators

- If we desire to estimate a parameter, we want to know that we are using a good estimator. We prefer for the estimator statistic to be an **unbiased** estimate of the parameter.
- A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ.
- If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$.

Standard Error

- The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$.

$X \quad \mu_X \quad \sigma_X$

$X_1, X_2, X_3, \ldots, X_n, \quad \bar{X}, \quad S$

- Examples of standard errors
  - $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.
  - $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.
- The problem is that often we do not know σ or p, for example. To get around this wee can use the estimators for these parameters. Then we have the **estimated standard error**.
- Again we need to know how these estimators $\hat{\theta}$ are being distributed.

What we use for estimating?

- A **confidence interval** is a range of possible values that is likely to contain the unknown population parameter we are seeking.
- First, we must have a **level of confidence**.
- Then based on this level we will compute a **margin of error**.
- Last, we can say that we are –% confident that the true population parameter falls within our confidence interval.

The confidence interval

The $1 - \alpha$ confidence interval for μ, given that we know the population standard deviation is:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Margin of Error

The margin of error is

$$m = \text{critical value} \times \text{standard error}$$

For means (given the population standard deviation is known), the margin of error is:

$$m = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

What is the margin of error for the mean monthly cell phone bill?

T distribution

- Used for the inference of the population mean. When population standard deviation σ is unknown.
- The distribution of the population is basically bell-shape.
- Formula for t:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$

- Use t-table, or qt(probability,df) in R.
- Degrees of freedom: $df = n - 1$.

Critical value when σ unknown

- When σ is **unknown** we use t-distribution.
- With degrees of freedom, $df = n - 1$.
- The critical value is $t_{\alpha/2}$ where the area between $-t_{\alpha/2}$ and $+t_{\alpha/2}$ under the T-curve is the confidence level $C = 1 - \alpha$.
- $t_{\alpha/2}$ is found in T-table using the row according to the degrees of freedom and the column according to the confidence level at the bottom of the table.
- In R use qt($(1 + C)/2$, df).

Confidence Interval for μ Recap

- Z-confidence interval, given the population standard deviation, σ is **known**

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- T-confidence interval, given that the population standard deviation, σ is **unknown**

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

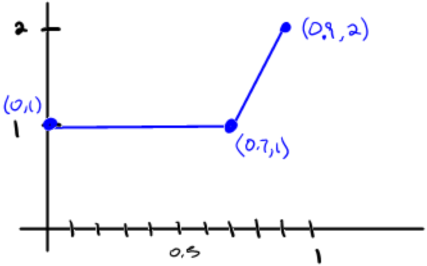
Choosing Sample Size

You can have both a high confidence while at the same time a small margin of error by taking enough observations.

- Sample size for confidence intervals of means.

$$n > \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$$

Think about a density curve that consists of two line segments. The first goes from the point (0, 1) to the point (0.7, 1). The second goes from (0.7, 1) to (0.9, 2) in the xy-plane. What percent of observations fall between 0.7 and 0.9?



$(0.7, 1) \to (0.9, 2)$

$$m = \frac{\Delta y}{\Delta x} = \frac{2-1}{0.9-0.7} = \frac{1}{0.2} = 5$$

Line: $y - 1 = 5(x - 0.7)$

$y = 5x - 3.5 + 1$

$y = 5x - 2.5$

The blue "curve" is graph of $f(x)$

What is $f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 0.7 \\ 5x - 2.5 & \text{if } 0.7 < x \le 0.9 \\ 0 & \text{otherwise} \end{cases}$

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 0.7 \\ 5x - 2.5 & \text{if } 0.7 < x \le 0.9 \\ 0 & \text{otherwise} \end{cases}$$

Is $1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_0^{0.7} 1\,dx + \int_{0.7}^{0.9} (5x - 2.5)\,dx$

$$= x \Big|_0^{0.7} + \left[\frac{5}{2}x^2 - 2.5x\right]_{0.7}^{0.9}$$

$$= 0.7 + \frac{5}{2}\left[0.9^2 - 0.9 - (0.7^2 - 0.7)\right]$$

$$= 0.7 + \frac{5}{2}\left[0.81 - 0.9 - 0.49 + 0.7\right]$$

$$= 0.7 + \frac{5}{2} \cdot 0.12 = 0.7 + 0.3 = 1$$

So, $P(0.7 \le x \le 0.9) = \int_{0.7}^{0.9} f(x)\,dx = \boxed{0.3}$ (see above)

Consider a spinner that, after a spin, will point to a number between zero and 1 with "uniform probability". Determine the probability: $P(1/9 \leq X \leq 23/45)$.

What is $f(x)$ here? $f(x) = \frac{1}{1-0} = 1$

$X \sim \text{Unif}(0,1)$

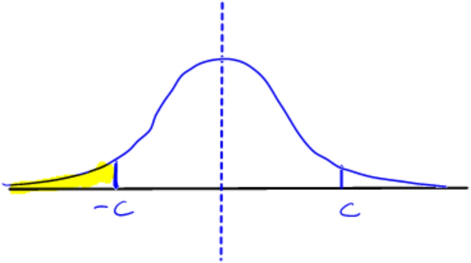
$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

$$P\left(\frac{1}{9} \leq X \leq \frac{23}{45}\right) = P\left(X \leq \frac{23}{45}\right) - P\left(X \leq \frac{1}{9}\right)$$

$$= F\left(\frac{23}{45}\right) - F\left(\frac{1}{9}\right)$$

$$= \frac{23}{45} - \frac{1}{9}$$

$$= \frac{18}{45} = \frac{2}{5} = 0.4$$



$$p = P(Z \le -c) = P(Z \ge c)$$

$$= 1 - P(Z \le c)$$

$$P(Z \le c) = 1 - P(Z \le -c)$$

$$c = \text{qnorm}(1 - p)$$

$$-c = -\text{qnorm}(1-p)$$

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/