Notes and Exercises Chapter 13 – Inference with Mathematical Models
Math 217
Saidi

Chapter 13 – Exploring the Normal Distribution

With randomization tests (Ch 11), the data were permuted assuming the null hypothesis. With bootstrapping (Ch 12), the data were resampled in order to measure the variability. In many cases the variability of the statistic can be described by the **computational method (as in previous chapters)** or by a **mathematical formula (as in this chapter).**

The **normal distribution** is presented here to describe the variability associated with sample proportions which are taken from either repeated samples or repeated experiments. The normal distribution is quite powerful in that it describes the variability of many different statistics, and we will encounter the normal distribution throughout the remainder of the book.

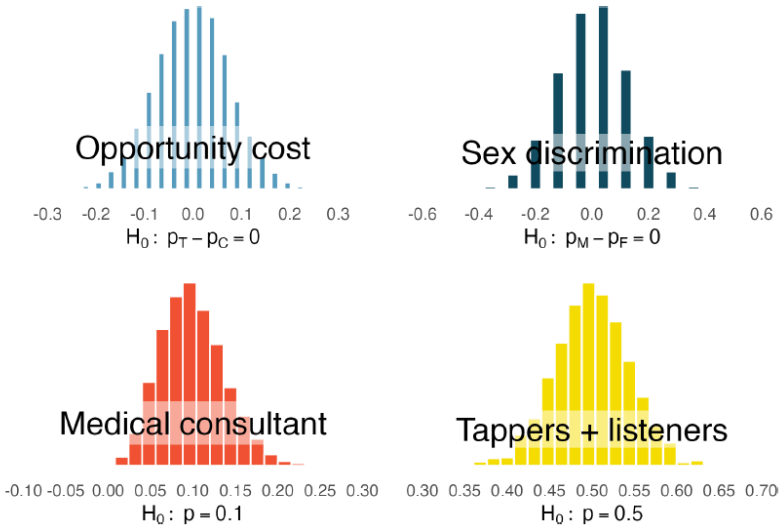
Focus on the parallels **between how data can provide insight about a research question** either **through computational methods** or **through mathematical models.**

13.1 Central Limit Theorem

Sampling distribution

A **sampling distribution** is the distribution of all possible values of a sample statistic from samples of a given sample size from a given population. We can think about the sample distribution as describing as how sample statistics (e.g. the sample proportion $\hat{p}$ or the sample mean $\bar{x}$) varies from one study to another. *A sampling distribution is contrasted with a data distribution which shows the variability of the observed data values.*

Figure 13.1 shows the null distributions in each of the four case studies where we ran 10,000 simulations. Note that the **null distribution** is the sampling distribution of the statistic created under the setting where the null hypothesis is true. Therefore, the null distribution will always be centered at the value of the parameter given by the null hypothesis. In the case of the opportunity cost study, which originally had just 1,000 simulations, we've included an additional 9,000 simulations.



The **null distribution** is the sampling distribution of the statistic created under the setting where the null hypothesis is true. Therefore, the null distribution will always be centered at the value of the parameter given by the null hypothesis.

Mathematical theory guarantees that if repeated samples are taken a sample proportion or a difference in sample proportions will follow something that resembles a normal distribution when certain conditions are met. These conditions fall into two general categories:

1. Observations in the sample are independent. Independence is guaranteed when we take a random sample from a population. Independence can also be guaranteed if we randomly divide individuals into treatment and control groups.
2. The sample is sufficiently large enough. What qualifies as "large enough" differs from one context to the next, and we'll provide suitable guidelines for proportions in Chapter 16. (For proportions, at least 10 expected successes and 10 expected failures in the sample)

The Normal Curve

$N(\mu = 0,\ \sigma = 1)$ $N(\mu = 19,\ \sigma = 4)$

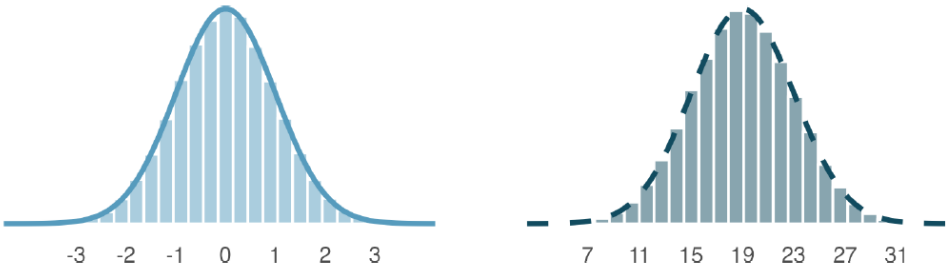


Figure 13.3: Both curves represent the normal distribution, however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 (blue solid line, on the left) is called the **standard normal distribution**. The other distribution (green dashed line, on the right) has mean 19 and standard deviation 4.

The normal distribution has notation: $Y \sim N(\mu,\sigma)$

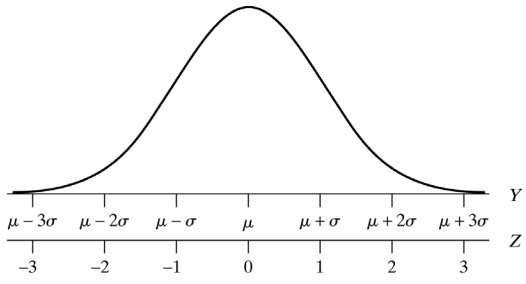
Y is the random variable of interest. We say "Random variable, Y, is approximately normally distributed with mean μ and standard deviation σ

Standard Normal Distribution has mean zero, standard deviation 1.

$Z \sim N(0,1)$ We read this as: Z is normally distributed with mean zero and standard deviation 1.

The Standardized Normal Curve

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



Example1 (using z-scores)

Suppose an admissions officer had 2 final candidates to consider for a spot at a university. Both were pretty much equally qualified in all senses, but one candidate took the ACT and the other took the SAT.

Since the ACT and SAT scores are on entirely different scales, the admissions officer needs a way to compare the two scores.

- Student A got 28 on the ACT, which has mean 21 and sd 5.4
- Student B got 1300 on the SAT, which has a mean of 1060 and sd 195.

Which student should be selected by the admissions officer? Calculate each z-score:

$$z_{ACT} =$$

$$z_{SAT} =$$

Conclusion:

13.2.3 Normal probability calculations

R-code with rpubs link: https://rpubs.com/rsaidi/1068797

Let q be an upper bound cutoff point on the z-scale or raw data scale.
Let p be a percentile (area shaded to the left)

Function	Description	Syntax	Example of use
PNORM	Cumulative Distribution Function (CDF)	pnorm(q, mean, sd)	pnorm(1.96, 0, 1) Gives the area under the standard normal curve to the left of 1.96, i.e. ~0.975
QNORM	Quantile Function – inverse of pnorm	qnorm(p, mean, sd)	qnorm(0.975, 0, 1) Gives the value at which the CDF of the standard normal is 0.975, i.e. ~1.96

Example2: Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm.

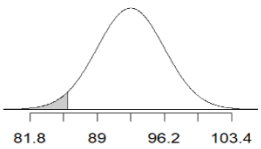
a. Compute the Z scores for possums with head lengths of 95.4 mm and 85.8 mm.
b. A brushtail possum is found to have a head length of 85.8 mm. What percentile of possums have head lengths 85.8 mm or smaller?
c. A brushtail possum is found to have a head length of 95.4 mm. What percentile of possums have head lengths 95.4 mm or more?

Answers

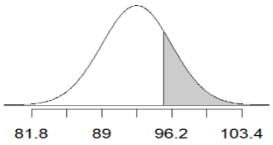
a. `z = (95.4-92.6)/3.6 = 0.7777778`
b. `pnorm(85.8, mean = 92.6, sd = 3.6) = 0.02945336`
c. `1 - pnorm(95.4, mean = 92.6, sd = 3.6) = 0.21835`
Alternatively, use: `pnorm(95.4, mean = 92.6, sd = 3.6, lower.tail = FALSE) = 0.21835`

To draw the pictures of these areas, use this code:

```
b. openintro::normTail(m = 92.6, s = 3.6, L = 85.8)
     (notice L is used to shade the lower tail)
```



c. `openintro::normTail(m = 92.6, s = 3.6, U = 95.4)`
(notice **U** is used to shade the upper tail)



Always draw a picture first, and find the Z score second.

For any normal probability situation, **always** draw and label the normal curve and shade the area of interest. The picture will provide an estimate of the probability.

After drawing a figure to represent the situation, identify the Z score for the observation of interest.

Using Percentiles to Calculate Raw Data Values

Example3: Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm.

a. What possum head length is in the bottom 20th percentile?
b. What head length is in the top 5th percentile?

Answers: (note **p** must be a **percentile,** which means it represents **area to the left of the cutoff)**

```
a. qnorm(p=.20, m=92.6, s=3.6) =  89.57016 mm
b. qnorm(p=.95, m=92.6, s=3.6) =  98.52147 mm
```

13.3.1 The Empirical Rule (68-95-99.7 Rule)

Typical Percentages: The Empirical Rule

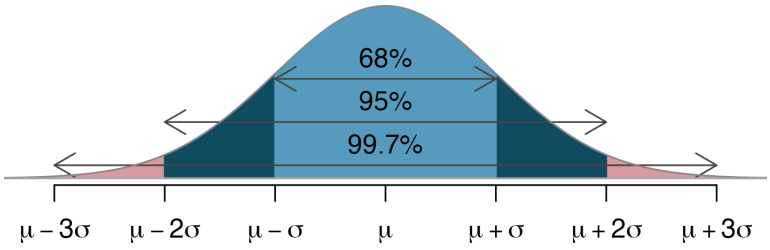
For "nicely shaped" distributions—that is, unimodal distributions that are not too skewed and whose tails are not overly long or short—we usually expect to find

about 68% of the observations within ±1 SD of the mean.
about 95% of the observations within ±2 SDs of the mean.
>99% of the observations within ±3 SDs of the mean.

Standard Deviation of the Sampling Distribution

We use the **mean of the sample** as the **mean of the sampling distribution.**

Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.



13.3.2 Standard error

The standard deviation of the sampling distribution is called the ***standard error.*** The **standard error** quantifies the variability of a point estimate from one sample to the next.

As n increases, precision increases and therefore spread decreases, which allows calculations to be more powerful at predicting the true parameter.

Whereas **standard deviations** describe the **dispersion** of observations, **standard errors** determine the **(un)reliability** of the statistics.

13.3.3 Margin of error for sample proportions

The distance given by z* × SE is called the margin of error.
Z* is the cutoff value found on the normal distribution. The most common value of z* is 1.96 (often approximated to be 2) indicating that the margin of error describes the variability associated with 95% of the sampled statistics.

13.4.3 Observed statistic vs. null statistics

As we learned in Section 13.2, it is helpful to draw and shade a picture of the normal distribution so we know precisely what we want to calculate. Here we want to find the area of the tail beyond 0.2, representing the p-value.

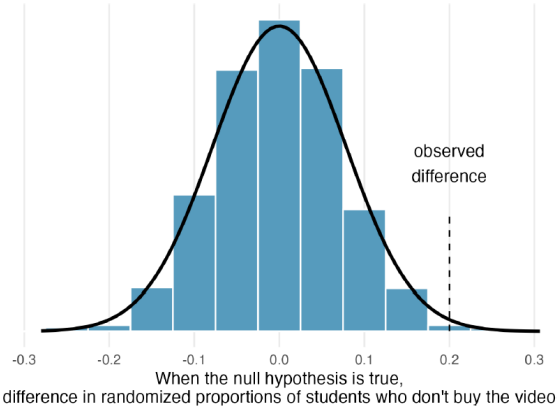


Figure 13.9: Null distribution of differences with an overlaid normal curve for the opportunity cost study. 10,000 simulations were run for this figure.

Z score in a hypothesis test

In the context of a hypothesis test, the Z score for a point estimate is: $z = \frac{statistic - null}{SE}$

Example4: Medical consultant

In Section 12.1 we learned about a medical consultant who reported that only 3 of their 62 clients who underwent a liver transplant had complications ($\hat{p} = \frac{3}{62} = 0.048$, which is less than the more common complication rate of 0.10. If we set the mean of 0.10 and borrow a formula that we'll encounter in Chapter 16, the standard error of this distribution was also computed at SE = 0.038. Find the z-score for the observed number of complications.

$$z_{complications} = \frac{statistic - null}{SE} = \frac{0.048 - 0.1}{0.038} = -1.37$$

This means that this consultant's clients experience complications at 1.37 standard deviations below the common rate. Later we will associate a p-value with this z-score.

13.5.4 Conditions for applying the normal model

Statistical techniques are like a carpenter's tools. When used responsibly, they can produce amazing and precise results. However, if the tools are applied irresponsibly or under inappropriate conditions, they will produce unreliable results. For this reason, with every statistical method that we introduce in future chapters, we will carefully outline conditions when the method can reasonably be used. These conditions should be checked in each application of the technique.

After covering the introductory topics in this course, advanced study may lead to working with complex models which, for example, bring together many variables with different variability structure. Working with data that come from normal populations makes higher-order models easier to estimate and interpret. There are times when simulation, randomization, or bootstrapping are unwieldy in either structure or computational demand. Normality can often lead to excellent approximations of the data using straightforward modeling techniques.

Confidence Intervals

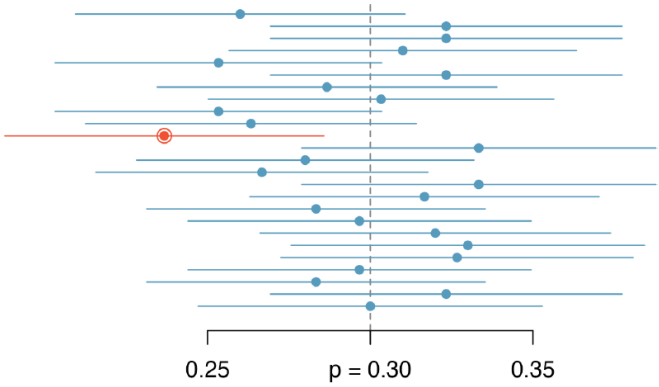


Figure 13.11: Twenty-five samples of size $n = 300$ were collected from a population with $p = 0.30$. For each sample, a confidence interval was created to try to capture the true proportion p. However, 1 of these 25 intervals did not capture $p = 0.30$.

The interval which does not capture p = 0.3 is not due to bad science. Instead, it is due to natural variability, and we should expect some of our intervals to miss the parameter of interest. Indeed, over a lifetime of creating 95% intervals, you should expect 5% of your reported intervals to miss the parameter of interest (unfortunately, you will not ever know which of your reported intervals captured the parameter and which missed the parameter).

Interpretation of 95% CI

We are 95% confident that the true population _____ (mean/proportion) for _____ (context of problem) is between _____ and _____.
(be sure to include units on your values)

Homework Chapter 13

1. Review section 13.7 (the chapter review)
2. **Suggested**: from textbook section 13.8 exercises: 1,3,5
3. **Required Lab:** Foundations for statistical inference – Confidence Intervals