19CSE304 - FOUNDATIONS OF DATA SCIENCE

FDS_03: Introduction to Data Science

Dr. Venugopal K

Assoc Professor, CSE, ASE, Amritapuri

Normal Distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the normal curve or normal distribution.

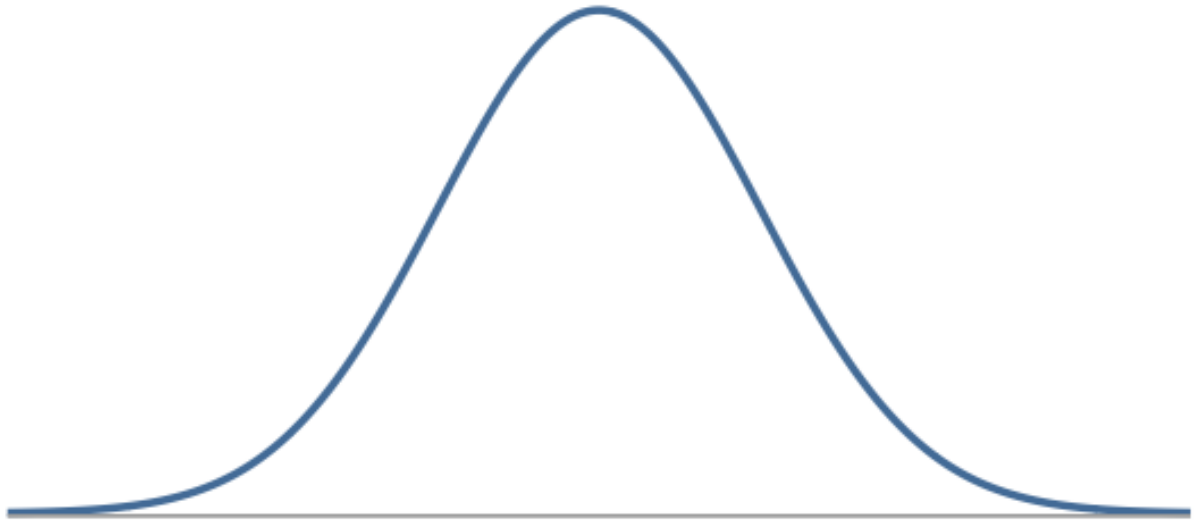
Variables such as SAT scores and heights of US adult males closely follow the normal distribution.

Normal Distribution

- Normal distribution
- Standardized scores
- Probabilities and percentiles
- 68-95-99.7% rule

Normal Distribution

- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma)$ → Normal with mean μ and standard deviation σ



Normal distribution model

- The normal distribution always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation.
- Changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve.

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**. The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**.

Figure A shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure B shows these distributions on the same axis.

Figure A

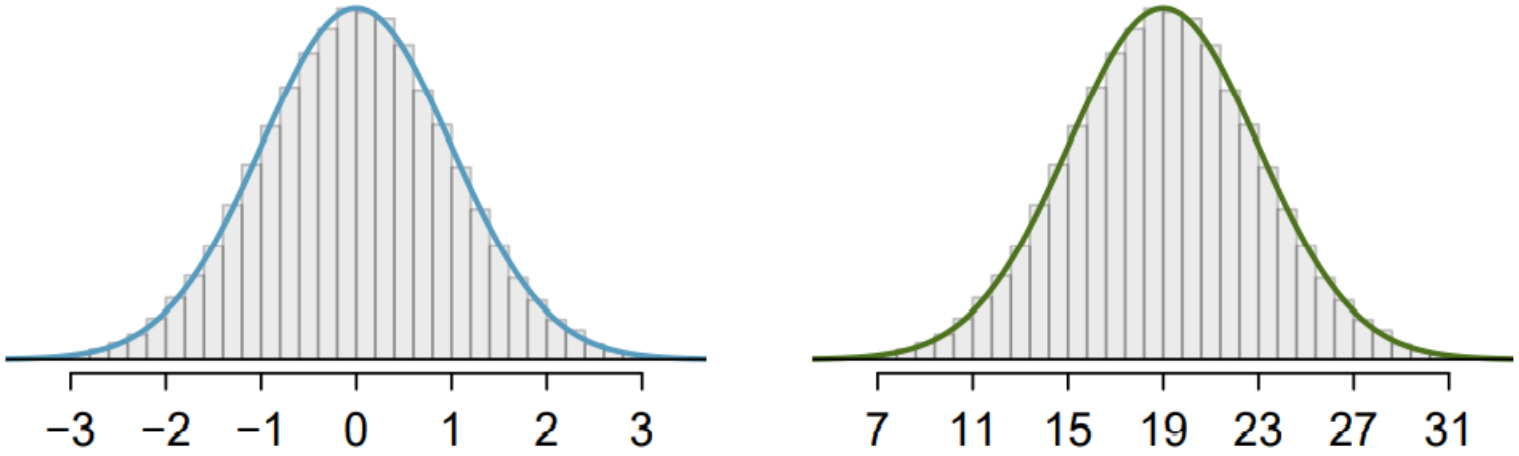
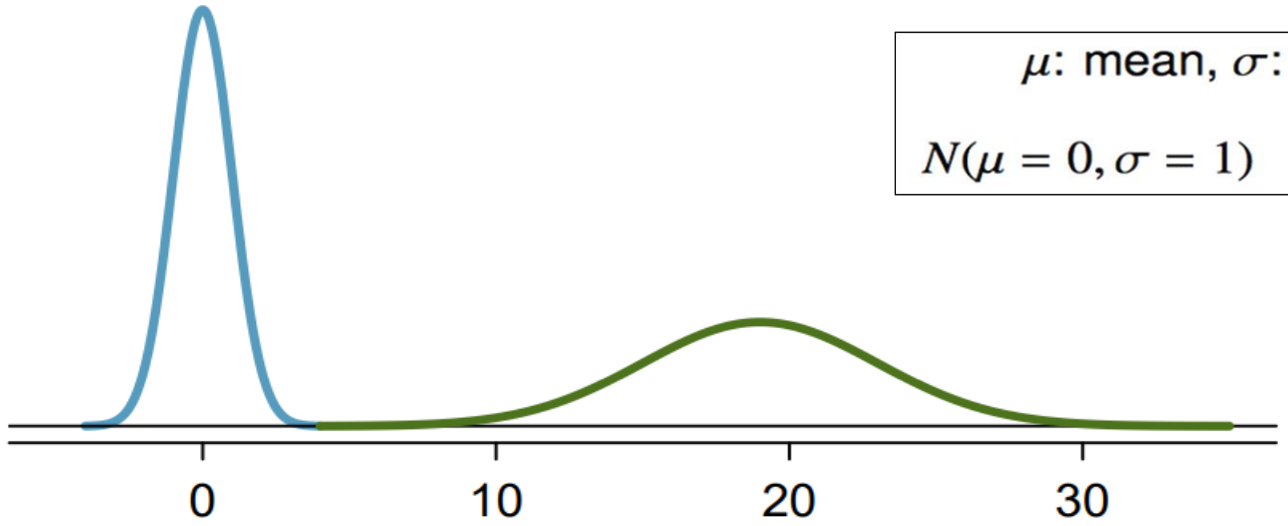


Figure B



μ: mean, σ: standard deviation

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

Standardizing with Z-scores

- We use a standardization technique called a Z-score, a method commonly employed for **nearly normal observations**, but may be used ***with any distribution***. *The Z-score of an observation is defined as the number of standard deviations it falls above or below the mean.*
- If the observation is one standard deviation above the mean, its Z-score is 1. If it is 1.5 standard deviations below the mean, then its Z-score is -1.5.
- If x is an observation from a distribution **N(μ, σ),** we define the Z-score mathematically as

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

Standardizing with Z scores (contd.)

These are called *standardized* scores, or *Z scores*.

- Z score of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{observation - mean}{SD}$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

Example 1

The following table shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

	SAT	ACT
Mean	1100	21
SD	200	6

Solution

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT:

1100 + 200 = 1300. Tom is 0.5 standard deviations above the mean on the ACT: 21 + 0.5 × 6 = 24.

	SAT	ACT
Mean	1100	21
SD	200	6

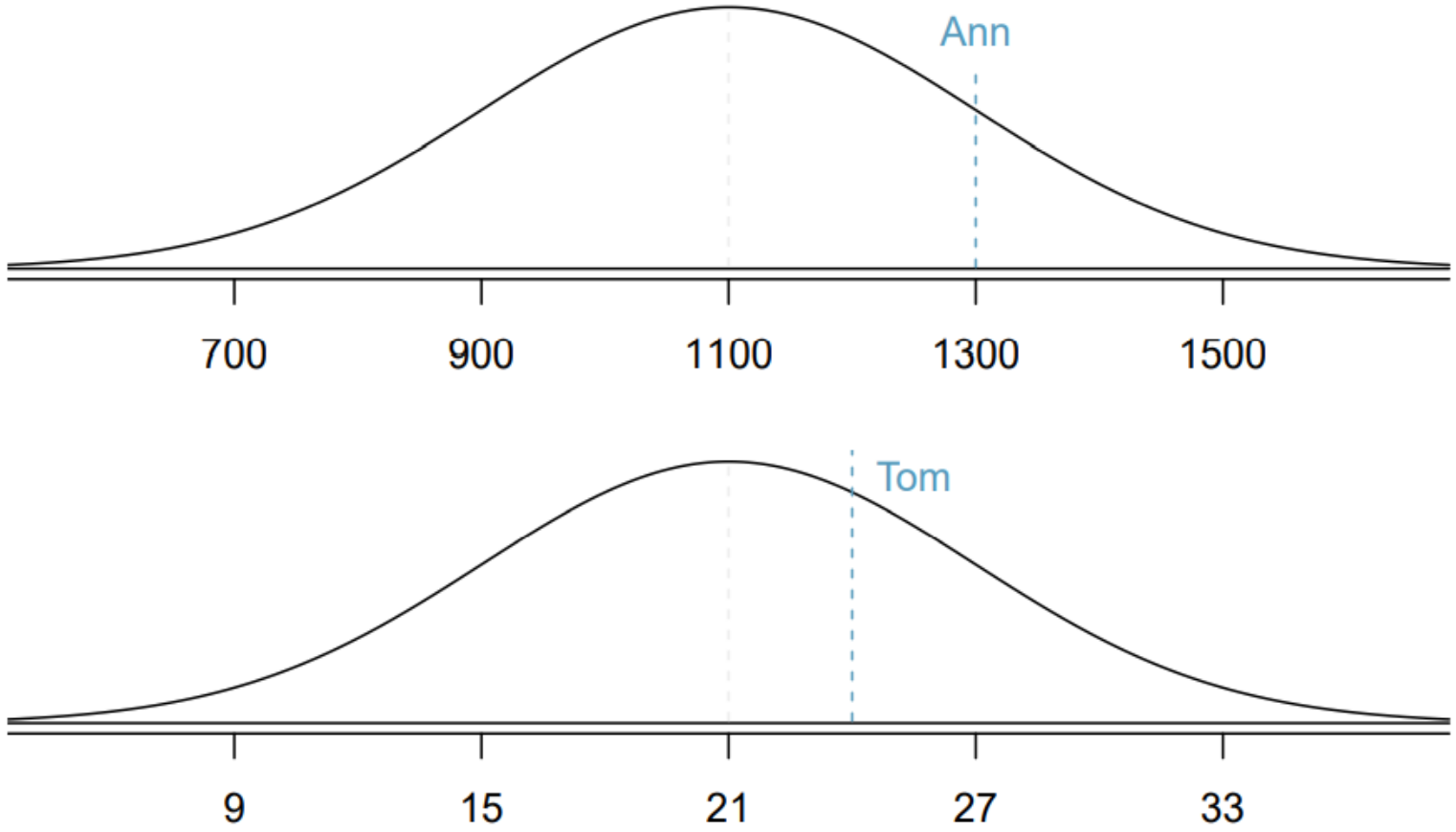
Calculation of Z-score

Using $\mu_{SAT} = 1100$, $\sigma_{SAT} = 200$, and $x_{\text{Ann}} = 1300$, we find Ann's Z-score:

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

Similarly for Tom, the Z-score can be calculated as:

$$Z_{Tom} = \frac{x_{\text{Tom}} - \mu_{\text{ACT}}}{\sigma_{\text{ACT}}} = \frac{24-21}{6} = 0.5$$

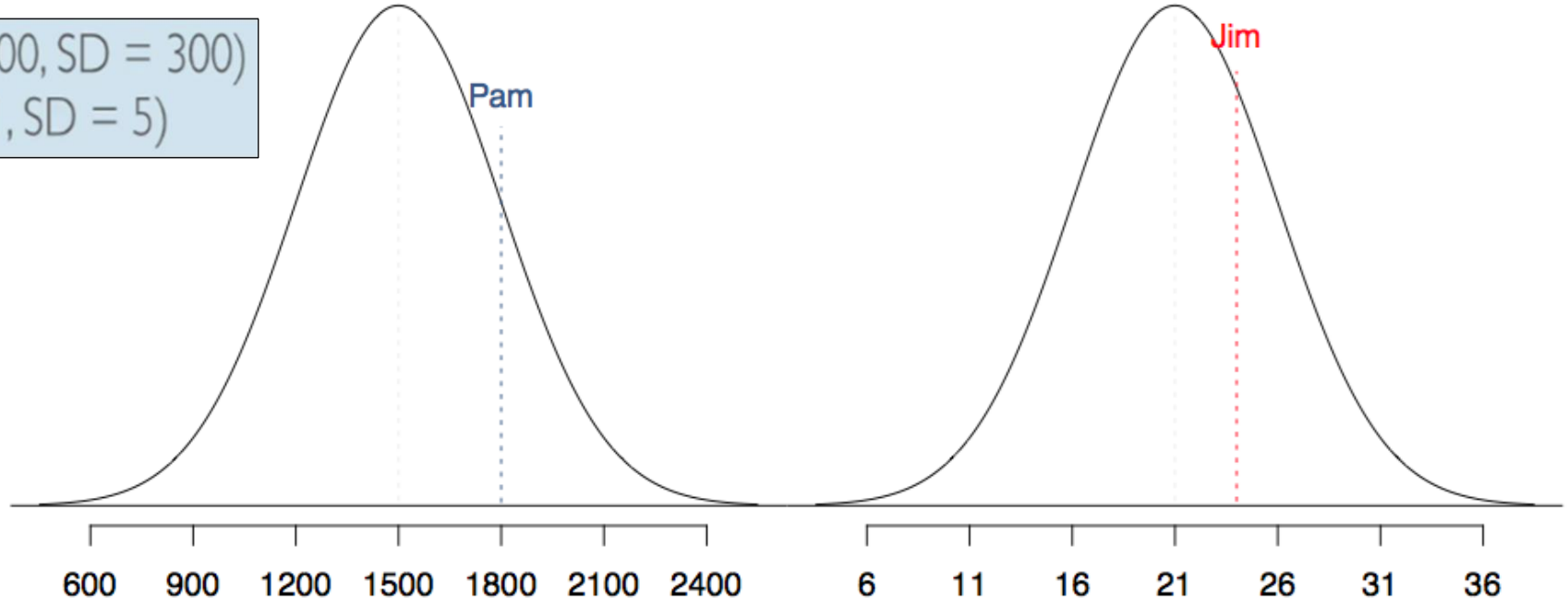


In the figure, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

Example 2

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

SAT scores ~ N(mean = 1500, SD = 300)
ACT scores ~ N(mean = 21, SD = 5)

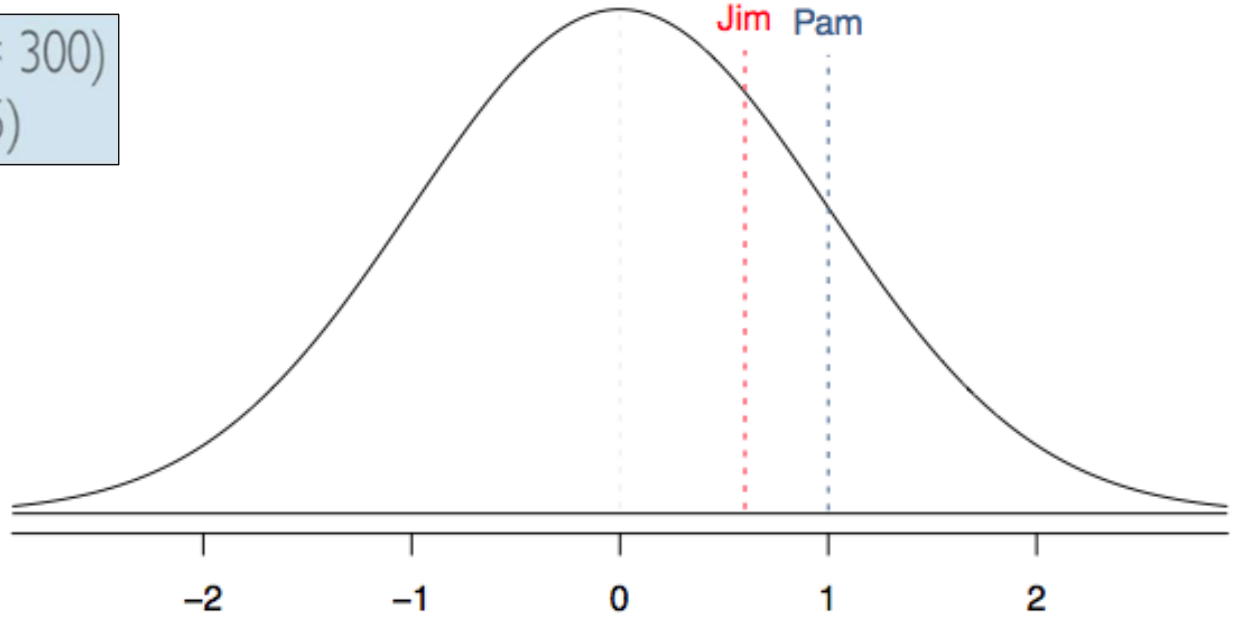


Example 2 – contd.

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is **(1800 - 1500) / 300 = 1** standard deviation above the mean.
- Jim's score is **(24 - 21) / 5 = 0.6** standard deviations above the mean.

SAT scores ~ N(mean = 1500, SD = 300)
ACT scores ~ N(mean = 21, SD = 5)



Example 3

Let X represent a random variable from **N(μ = 3, σ = 2)**, and suppose we observe **x = 5.19**.

(a) Find the Z-score of x.

(b) Use the Z-score to determine how many standard deviations above or below the mean x falls.

Solution:

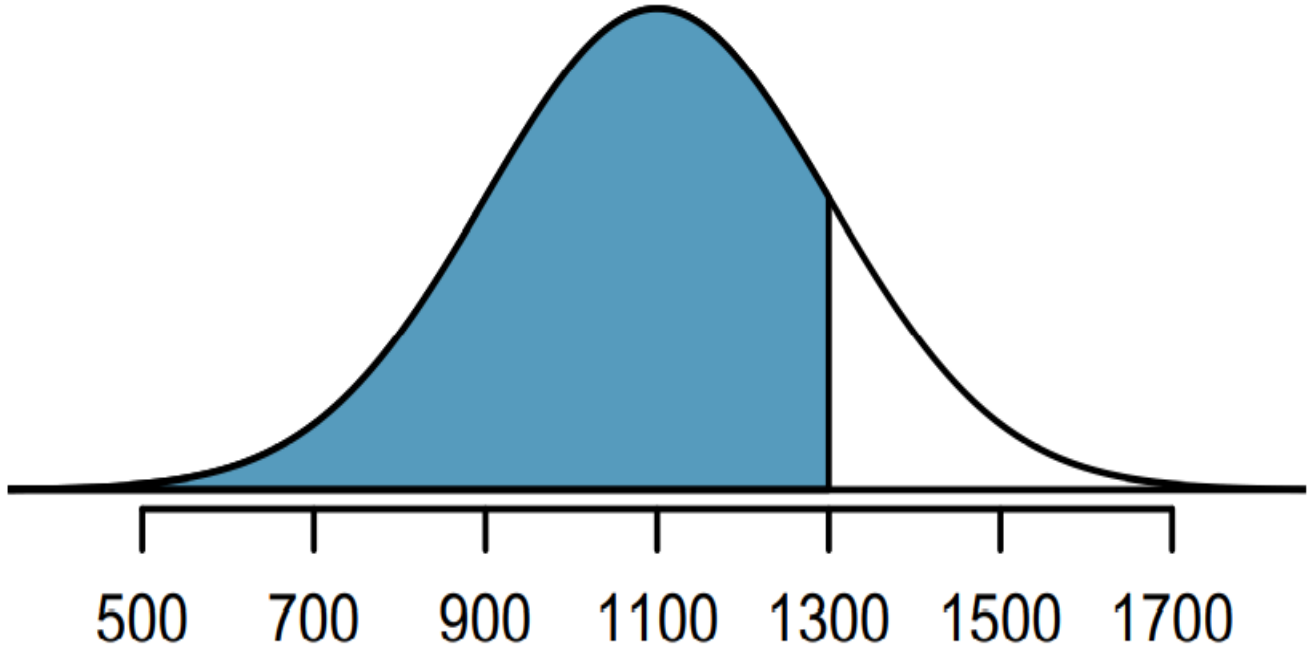
(a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$

(b) The observation x is 1.095 standard deviations above the mean. We know it must be above the mean since Z is positive.

Finding tail areas(Percentiles)

- *Percentile* is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of that observation.
- For instance, what fraction of people have an SAT score below Ann's score of 1300? This is the same as the percentile Ann is at, which is the percentage of cases that have lower scores than Ann. We can visualize such a tail area like the curve and shading shown in the figure below.

	SAT	ACT
Mean	1100	21
SD	200	6



Calculating percentiles – (1) using computation

There are different ways to compute percentiles/areas under the curve.

(1) In R type in the following command in the console window:

```
> pnorm(1300, mean = 1100, sd = 200)
[1] 0.8413447
```

According to this calculation, the region shaded that is below 1300 represents the proportion 0.841 (84.1%) of SAT test takers who had SAT score below 1300, or equivalently Z-scores below Z = 1.

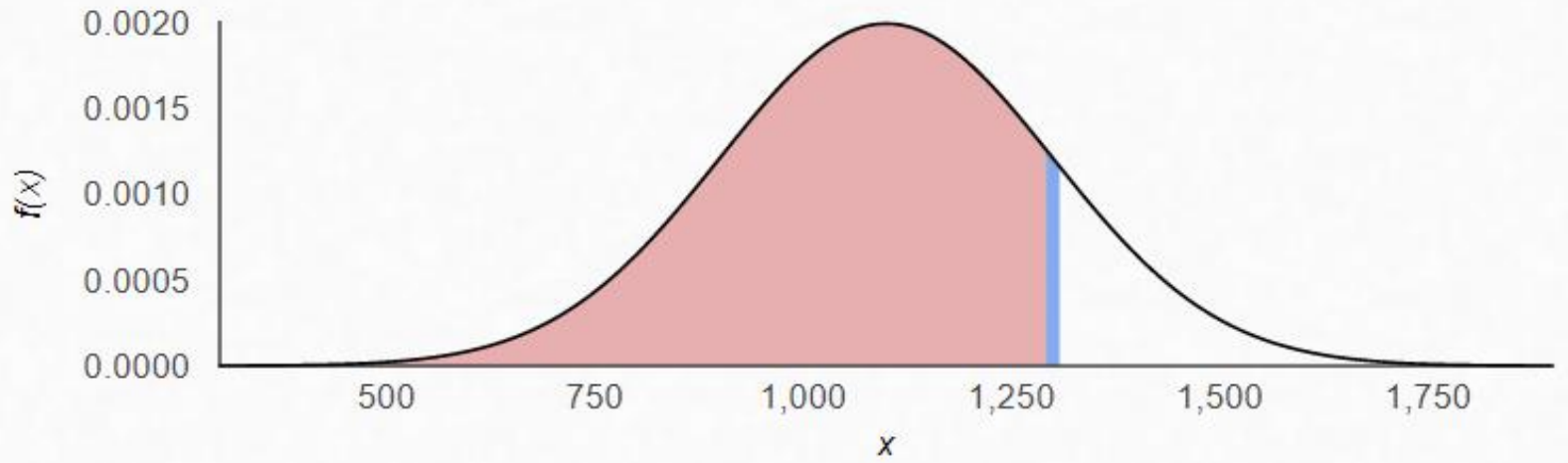
(2) URL: [Normal Distribution Applet/Calculator (uiowa.edu)]

Normal Distribution

$$X \sim N(\mu, \sigma)$$

$\mu =$ 1100 $\quad \sigma =$ 200

$x =$ 1300 $\quad$ P(X < x) = $\quad$ 0.84134



$$\mu = E(X) = 1100 \quad \sigma = SD(X) = 200 \quad \sigma^2 = Var(X) = 40000$$

(3) Calculating percentiles - using tables

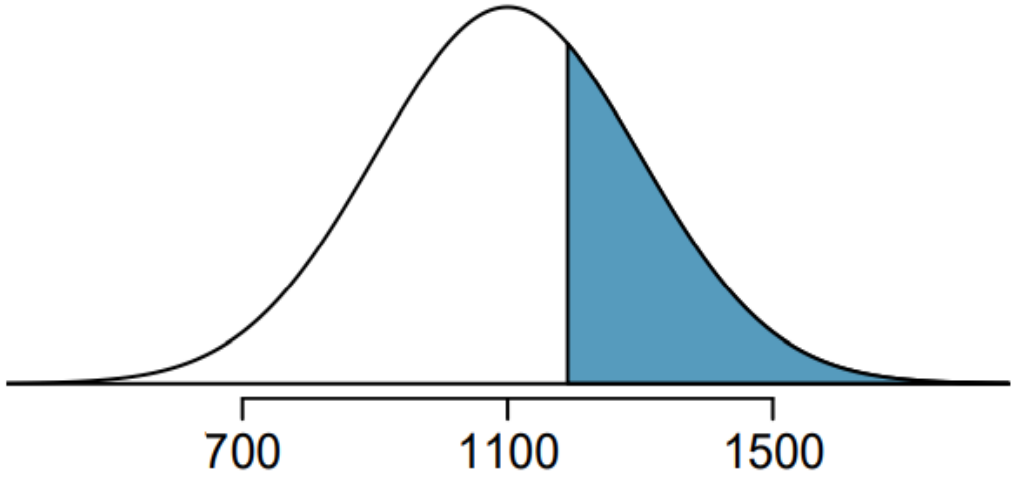
URL: [*Microsoft Word - STU Z Table.doc (arizona.edu)]

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Example 4

Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1190 on her SATs?

Solution: First, always draw and label a picture of the normal distribution. We are interested in the chance she scores above 1190, so we shade this upper tail:



Note: Cumulative SAT scores are approximated well by a normal model, **N(μ = 1100, σ = 200).**

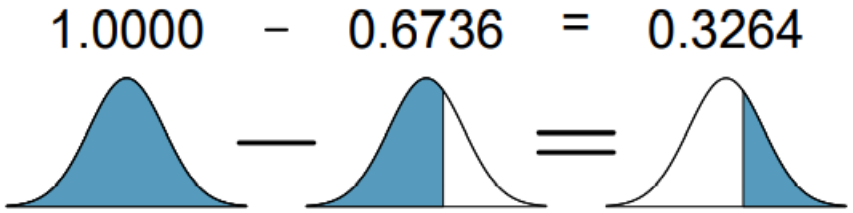
Example 4 – contd.

To find the shaded area under the curve make use of the Z-score of the cutoff value. With **μ = 1100, σ = 200**, and the cutoff value **x = 1190**, the

Z-score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

Using statistical software, or another method, we can area left of **Z = 0.45** as **0.6736**. To find the area above Z = 0.45, we compute one minus the area of the lower tail

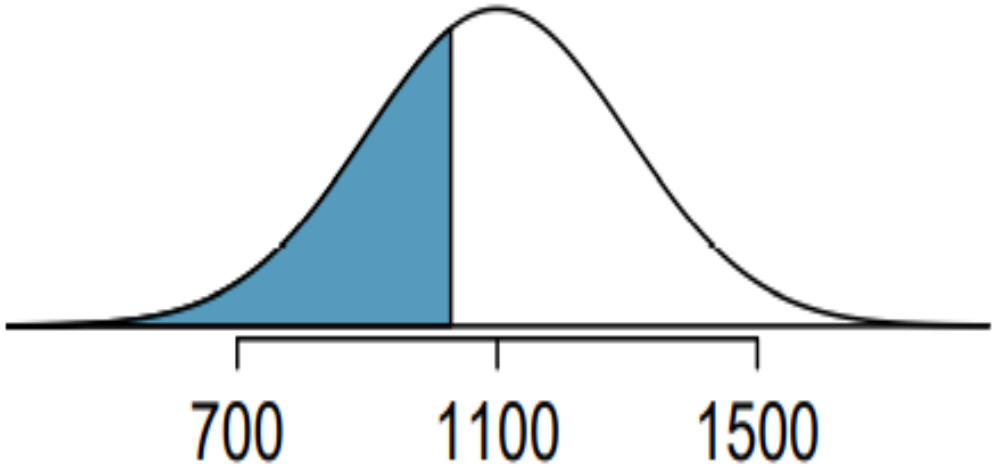


The probability Shannon scores at least 1190 on the SAT is **0.3264**.

Example 5

Edward earned a 1030 on his SAT. What is his percentile?

First, a picture is needed. ***Edward's percentile is the proportion of people who do not get as high as a 1030. These are the scores to the left of 1030.***



Note: Cumulative SAT scores are approximated well by a normal model, **N(μ = 1100, σ = 200).**

Example 5 – contd.

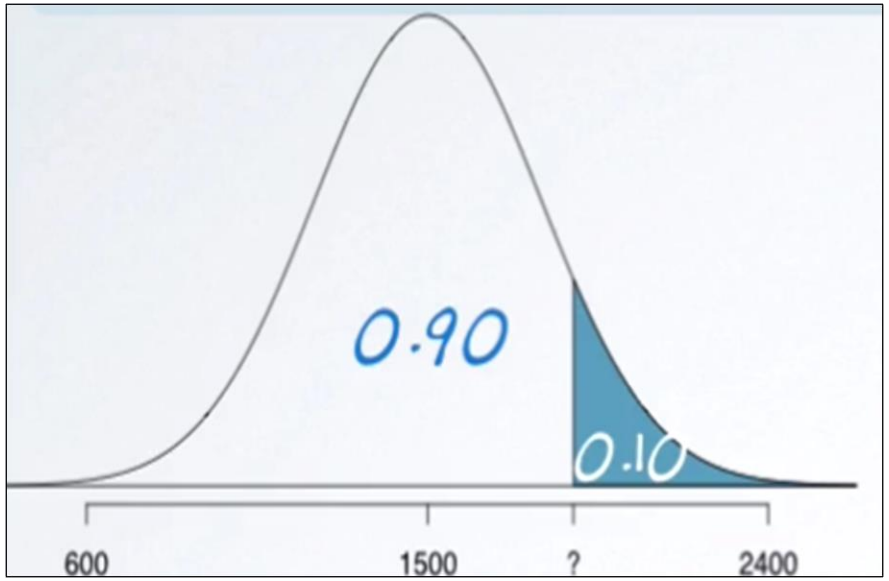
Identifying the mean μ = 1100, the standard deviation σ = 200, and the cutoff for the tail area x = 1030 makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Using statistical software, we get a tail area of 0.3632. Edward is at the 36th percentile.

Example 6

A friend informed you that she scored in the top 10% of SAT. What is the lowest possible score that she could have gotten?



$$Z = (X - 1500)/300$$

$$X = (1.28 \times 300) + 1500$$

$$= 1884$$

z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

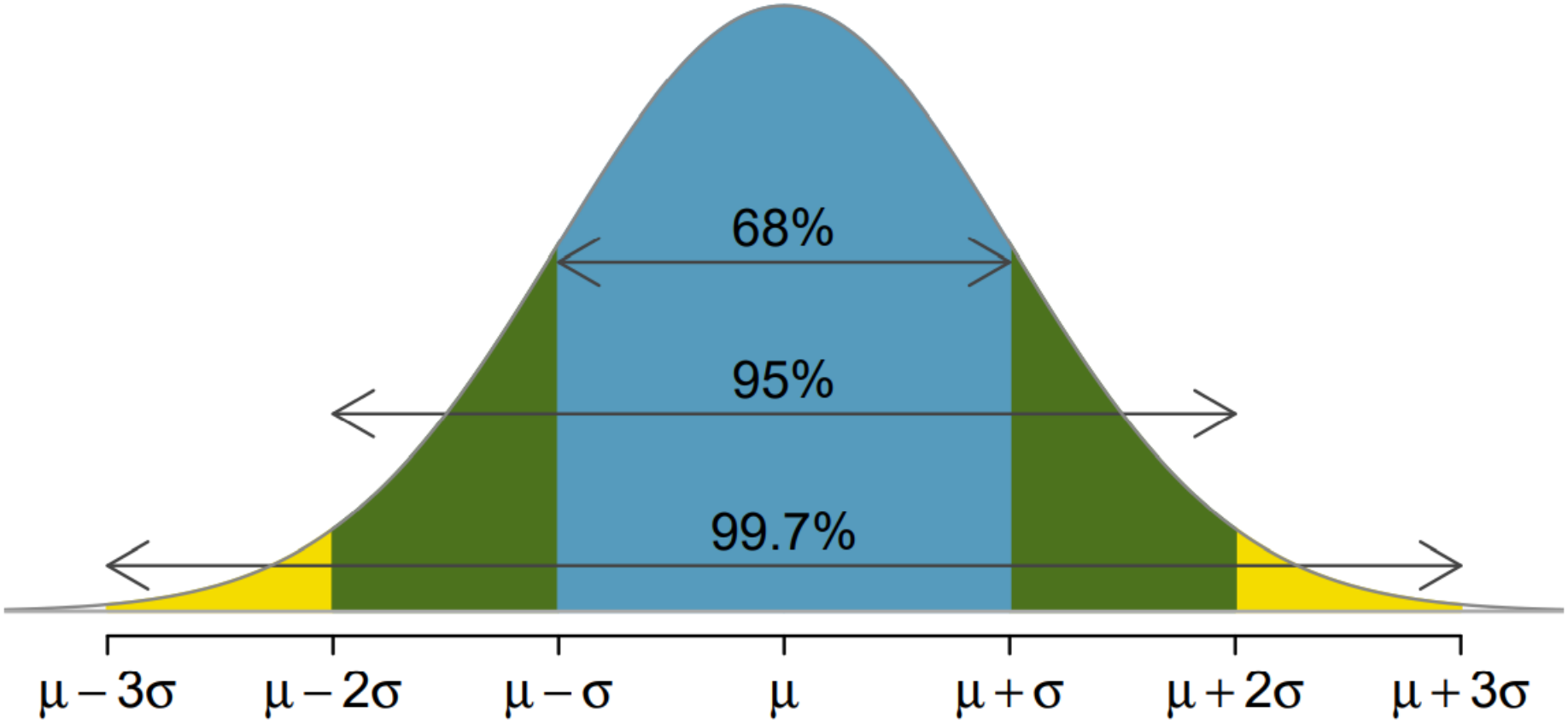
Example 6 – contd.

Alternately using a computational approach:

```
R
> qnorm(0.90, 1500, 300)
[1] 1884.465
```

68-95-99.7% rule

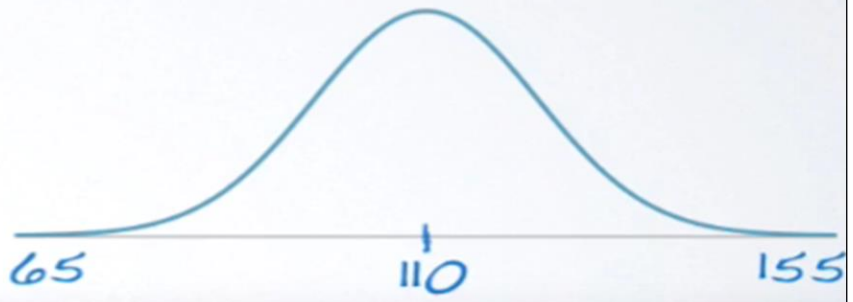
There is a useful rule of thumb that about **68%**, **95%**, and **99.7%** of observations fall within **1**, **2**, and **3**, standard deviations respectively of the mean in a normal distribution.



Example

A doctor collects a large set of heart rate measurements that approximately follow a normal distribution. He only reports 3 statistics, the mean = 110 beats per minute, the minimum = 65 beats per minute, and the maximum = 155 beats per minute. Which of the following is most likely to be the standard deviation of the distribution?

(a) 5
(b) 15
(c) 35
(d) 90



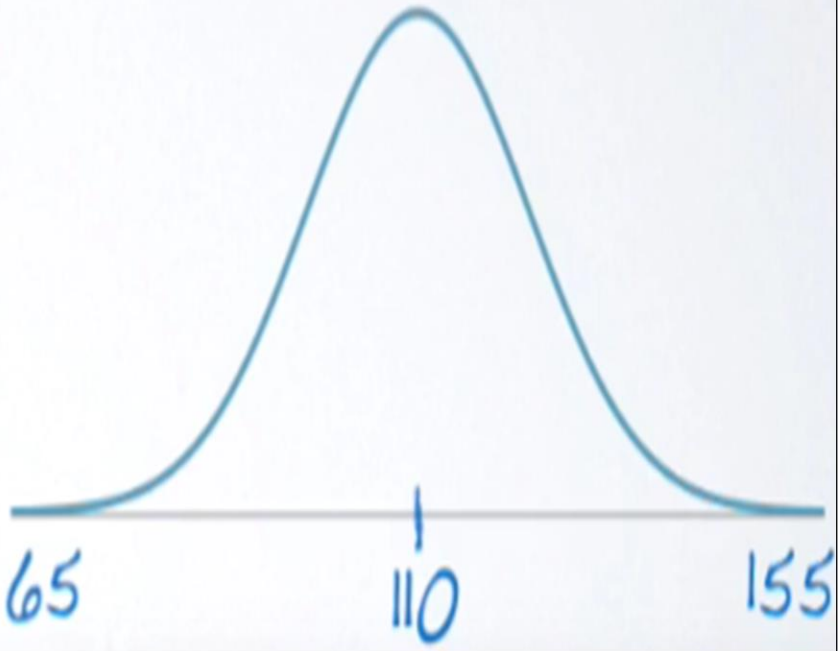
Example – contd.

(a) 5 → $110 \pm (3 \times 5) = (95, 125)$

(b) 15 → $110 \pm (3 \times 15) = (65, 155)$

(c) 35 → $110 \pm (3 \times 35) = (5, 215)$

(d) 90 → $110 \pm (3 \times 90) = (-160, 380)$

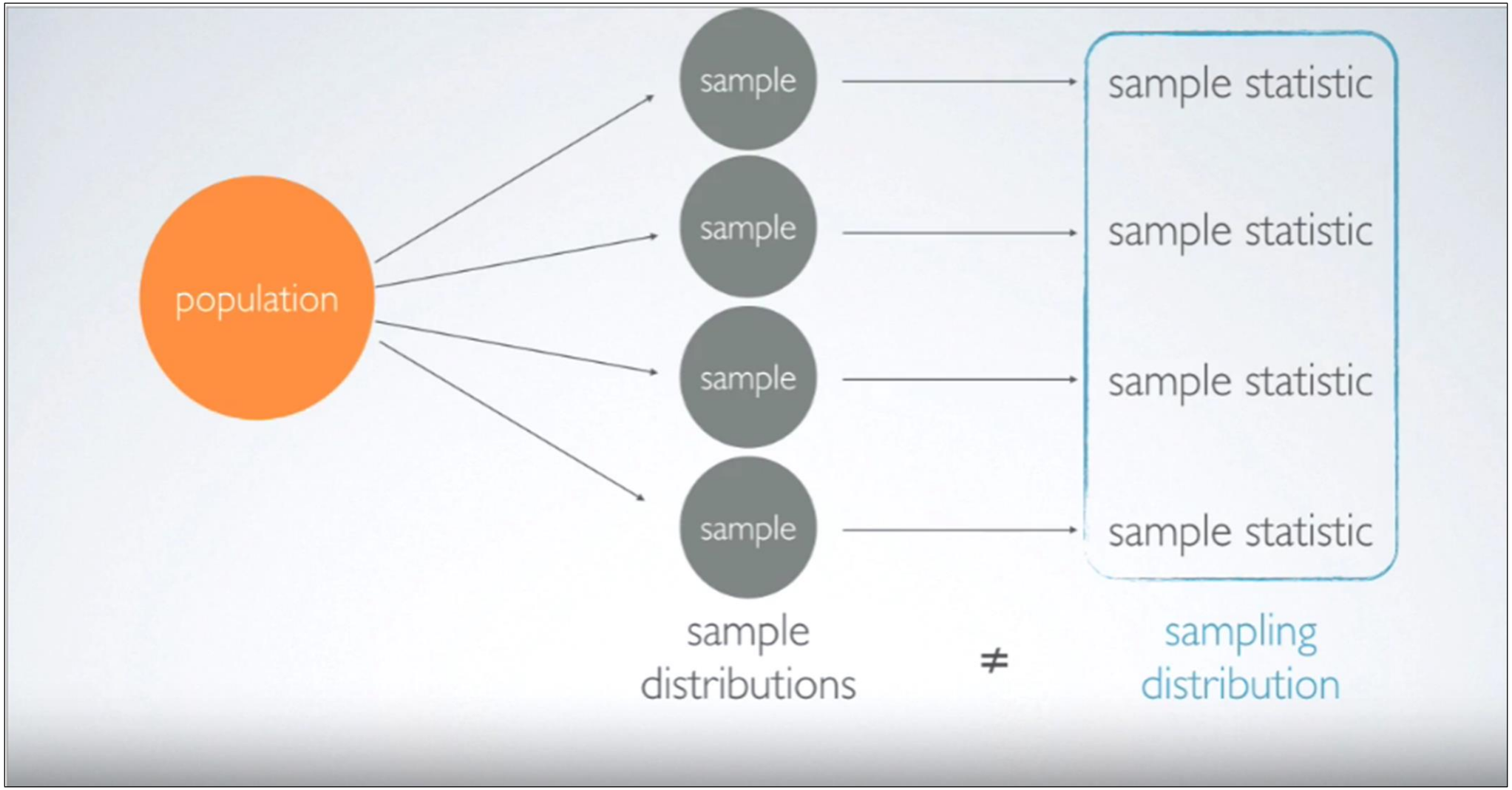


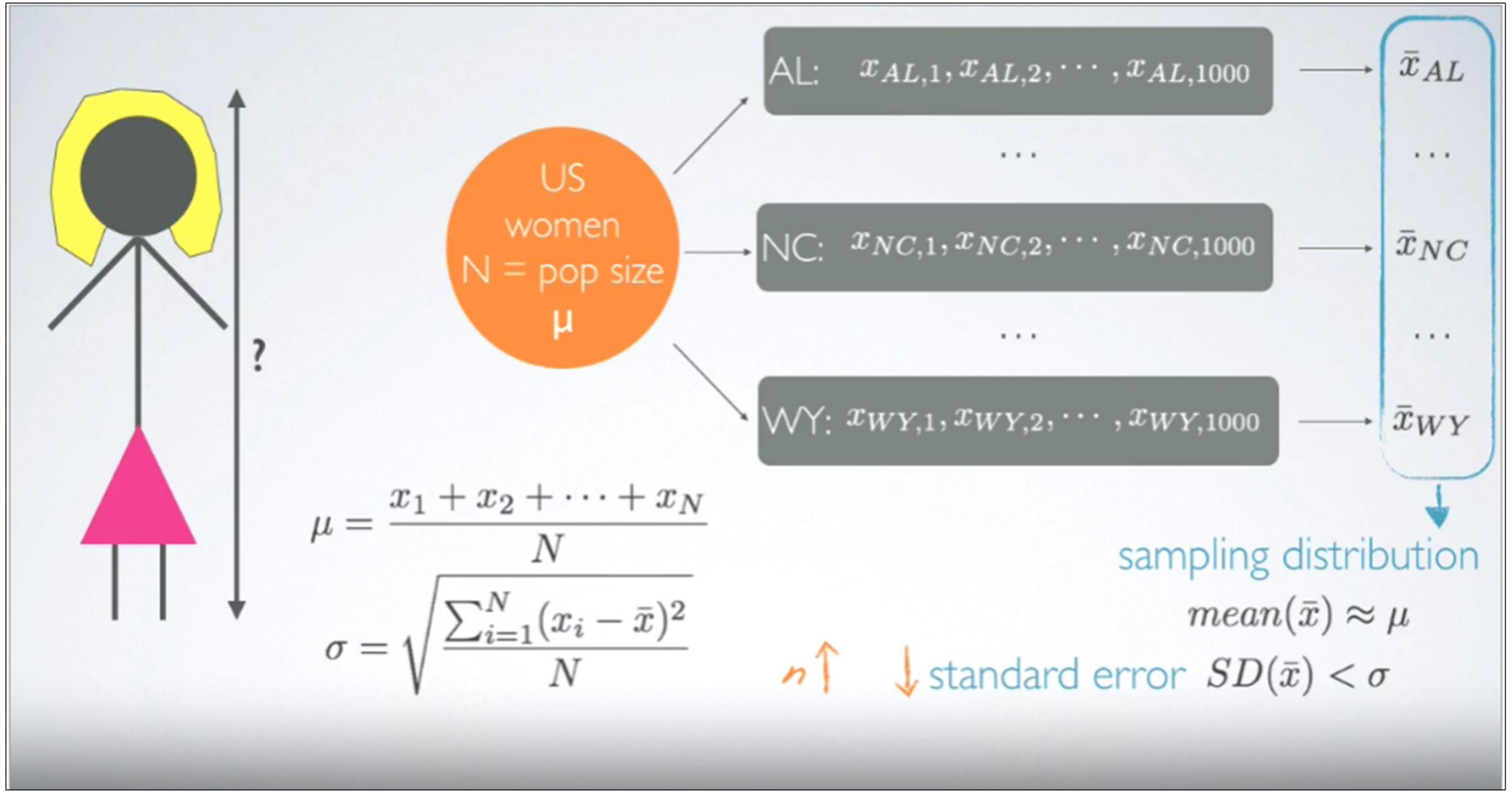
Basics of Inference – Point Estimates

Point Estimates

Topics

- Point estimates → what they are, and how we use them.
- Sampling error and bias.
- Introduce the concepts of a sampling distribution of a point estimate, and discuss its properties.
- Central Limit Theorem → when it applies to a sample proportion, and what happens when it doesn't apply.
- How these same ideas apply in many other applications.


population
sample
sample
sample
sample
sample statistic
sample statistic
sample statistic
sample statistic
sample distributions
≠
sampling distribution


?
US
women
N = pop size
μ
AL: $x_{AL,1}, x_{AL,2}, \cdots, x_{AL,1000}$
NC: $x_{NC,1}, x_{NC,2}, \cdots, x_{NC,1000}$
WY: $x_{WY,1}, x_{WY,2}, \cdots, x_{WY,1000}$
$\bar{x}_{AL}$
$\bar{x}_{NC}$
$\bar{x}_{WY}$
$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$
$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$
sampling distribution
$mean(\bar{x}) \approx \mu$
n ↑ ↓ standard error $SD(\bar{x}) < \sigma$

Example – President's approval rating

Suppose we take a random sample of 780 American adults and find that 45% of them approve of the job done by the US President.

While the estimate might not be perfect, we would expect that 45% is a pretty good estimate of the approval rating.

But does this mean that exactly 45% of *all* American adults approve of the job done by the US President?

As stated, a poll suggested the US President's approval rating is 45%. We would consider 45% to be a **point estimate** of the approval rating we might see if we collected responses from the entire population. This **entire-population response proportion** is generally referred to as the **parameter of interest**. When the parameter is a proportion, it is often denoted by **p**, and we often refer to the sample proportion as $\hat{p}$. Unless we collect responses from every individual in the population, p remains unknown, and we use $\hat{p}$ as our estimate of **p**. The difference we observe from the poll versus the parameter is called the error in the estimate.

Generally, the error consists of two aspects: **sampling error** and **bias**.

Errors in point estimates

There are two main categories of uncertainty in point estimates.

Sampling Error

The error resulting from the randomness in the sampling process.

Much of statistics is focused on sampling error.

Bias

Systematic tendency to over- or under-estimate the parameter.

Critical to keep in mind, especially during data collection

Sampling error

Sampling error, sometimes called sampling uncertainty, describes how much an estimate will tend to vary from one sample to the next. For instance, the estimate from one sample might be 1% too low while in another it may be 3% too high. Much of statistics is focused on understanding and quantifying sampling error, and we will find it useful to consider a sample's size to help us quantify this error; the sample size is often represented by the letter **n**.

Bias

Bias describes a systematic tendency to over- or under-estimate the true population value.

For example, if we were taking a student poll asking about support for a new college stadium, we'd probably get a biased estimate of the stadium's level of student support by wording the question as, *Do you support your school by supporting funding for the new stadium?*

We try to minimize bias through thoughtful data collection procedures as described earlier.

Understanding the variability of a point estimate

Suppose the proportion of American adults who support the expansion of solar energy is **p = 0.88**, which is our parameter of interest.

If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%? We want to understand, how does the sample proportion $\hat{p}$ behave when the true population proportion is 0.88.

Understanding the variability of a point estimate

We can simulate responses we would get from a simple random sample of 1000 American adults, which is only possible because we know the actual support for expanding solar energy is 0.88. Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write "support" on 88% of them and "not" on the other 12%.

2. Mix up the pieces of paper and pull out 1000 pieces to represent our sample of 1000 American adults.

3. Compute the fraction of the sample that say "support".

Understanding the variability of a point estimate

Running this simulation with 250 million pieces of paper would be time-consuming and very costly, but we can simulate it using computer code.

In the simulation, the sample gave a point estimate of $\hat{p}_1$ = 0.886. We know the population proportion for the simulation was **p = 0.88**, so we know the estimate had an error of 0.886 − 0.88 = +0.006.

One simulation isn't enough to get a great sense of the distribution of estimates we might expect in the simulation, so we should run more simulations. In a second simulation, we get $\hat{p}_2$ = 0.887, which has an error of +0.007. In another, $\hat{p}_3$ = 0.884 for an error of +0.004.

Understanding the variability of a point estimate

With the help of a computer, we've run the simulation 10,000 times and created a histogram of the results from all 10,000 simulations in the following figure. This distribution of sample proportions is called a sampling distribution. We can characterize this sampling distribution as follows:

Center. The center of the distribution is $\bar{x}_{\hat{p}} = 0.880$, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straightforward sampling strategy that helps avoid sampling bias.

Spread. The standard deviation of the distribution is $s_{\hat{p}} = 0.010$. When we're talking about a sampling distribution or the variability of a point estimate, we typically use the term **standard error** rather than *standard deviation*, and the notation $SE_{\hat{p}}$ is used for the standard error associated with the sample proportion.

Shape. The distribution is symmetric and bell-shaped, and it *resembles a normal distribution.*

R-code for generating Point Estimate, and plot of Sample Distribution

```
# Create a set of 250 million entries, where 88% of them are "support“, and 12% are "not".
pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size), rep("not", 0.12 * pop_size))

# Function for sampling, and computing phat.
phat <- function(x = possible_entries){
  sampled_entries <- sample(x, size = 1000)
  return(sum(sampled_entries == "support") / 1000)
}

# Execute function ten thousand times to populate vector phat_10k
phat_10k <- replicate(10000, phat())
```

```
> head(phat_10k)
[1] 0.886 0.887 0.884 0.883 0.873 0.875
```

```
# Plot the histogram of Sample Proportions
hist(phat_10k, main="Sample Distribution", xlab="Sample Proportions",
ylab="Frequency", border="black", col="blue")
```

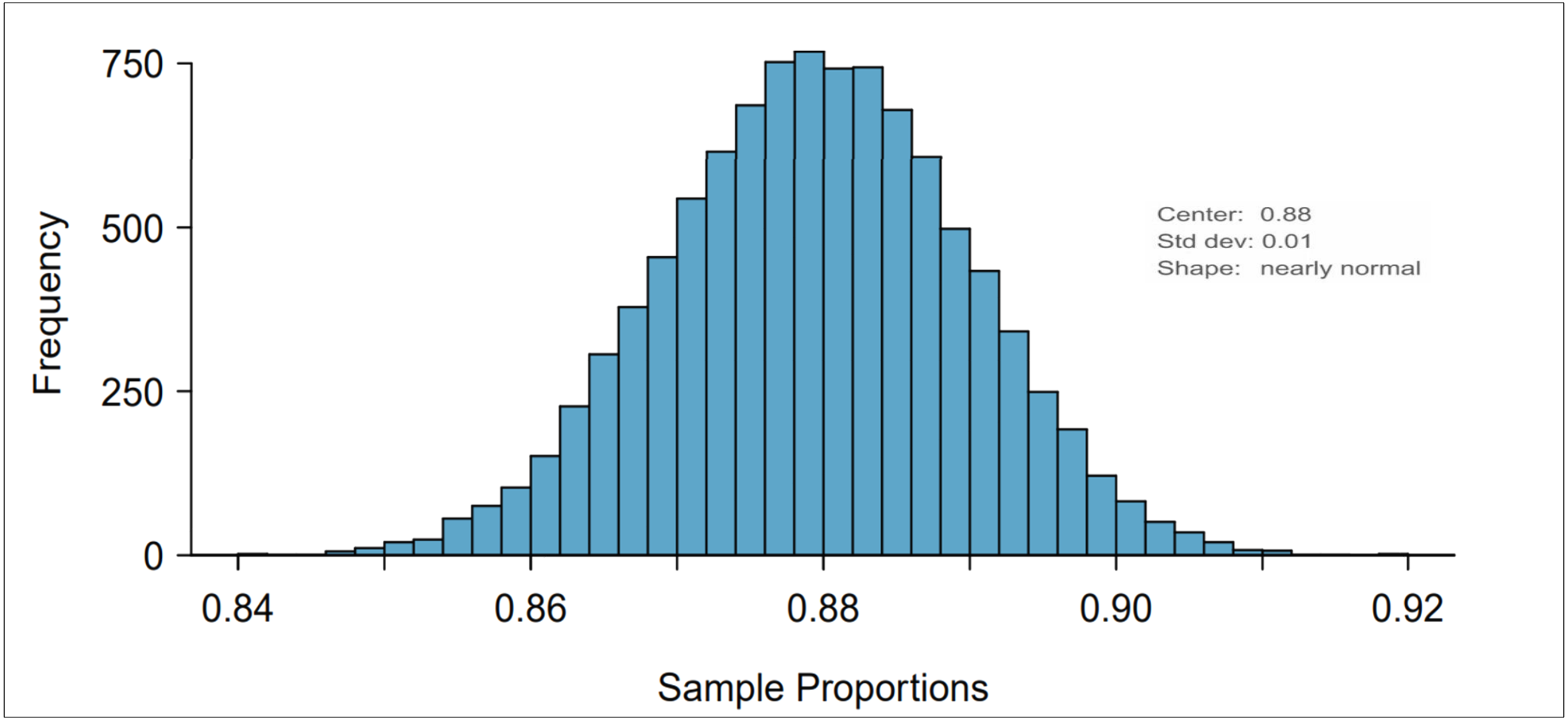


Fig: A histogram of 10,000 sample proportions, where each sample is taken from a population where the population proportion is 0.88 and the sample size is n = 1000.

Understanding the variability of a point estimate

Note: When the population proportion is **p = 0.88** and the sample size is n = 1000, the sample proportion $\hat{p}$ tends to give a pretty good estimate of the population proportion. We also have the interesting observation that the histogram resembles a normal distribution.

Central Limit Theorem

The distribution in the above figure looks a lot like a normal distribution. That is no anomaly; it is the result of a general principle called the Central Limit Theorem.

CENTRAL LIMIT THEOREM AND THE SUCCESS-FAILURE CONDITION

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.

URL: [CLT for means (shinyapps.io)]

CLT – conditions

Certain conditions must be met for the CLT to apply:

Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if

- random sampling / assignment is used, and
- if sampling without replacement, n < 10% of the population.

Sample size / skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.

- the more skewed the population distribution, the larger sample size we need for the CLT to apply
- for moderately skewed distributions n > 30 is a widely used rule of thumb

This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.

Central Limit Theorem

The Central Limit Theorem is incredibly important, and it provides a foundation for much of statistics. As we begin applying the Central Limit Theorem, be mindful of the two technical conditions: the observations must be independent, and the sample size must be sufficiently large such that $np \geq 10$ and $n(1-p) \geq 10$.

Earlier we estimated the mean and standard error of $\hat{p}$ using simulated data when **p = 0.88** and **n = 1000**. Confirm that the Central Limit Theorem applies and the sampling distribution is approximately normal.

Central Limit Theorem

Independence. There are $n = 1000$ observations for each sample proportion $\hat{p}$, and each of those observations are independent draws. *The most common way for observations to be considered independent is if they are from a simple random sample.*

Success-failure condition. We can confirm the sample size is sufficiently large by checking the success-failure condition and confirming the two calculated values are greater than 10:

$$np = 1000 \times 0.88 = 880 \geq 10 \qquad n(1-p) = 1000 \times (1 - 0.88) = 120 \geq 10$$

The independence and success-failure conditions are both satisfied, so the Central Limit Theorem applies, and it's reasonable to model $\hat{p}$ using a normal distribution.

Example

Compute the theoretical mean and standard error of $\hat{p}$ when $p = 0.88$ and $n = 1000$, according to the Central Limit Theorem.

The mean of the $\hat{p}$'s is simply the population proportion: $\mu_{\hat{p}} = 0.88$.

The calculation of the standard error of $\hat{p}$ uses the following formula:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88(1-0.88)}{1000}} = 0.010$$

Closing notes on CLT and Sampling

- Sample statistic is a point estimate for a population parameter, for example, the sample mean is used to estimate the population mean. Note that point estimate and sample statistic are synonymous.
- Point estimates (such as the sample mean) will vary from one sample to another, and this variability is the sampling variability.
- We need to distinguish standard deviation (**σ or s**) and standard error (**SE**): standard deviation measures the variability in the data, while standard error measures the variability in point estimates from different samples of the same size and from the same population, i.e. measures the sampling variability.
- Conceptually: Imagine taking many samples from the population. When the size of each sample is large, the sample means will be much more consistent across samples than when the sample sizes are small.
- Remember $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. When the sampling size *n* increases, the sampling variability will decrease since n is in the denominator

Confidence Intervals for a Proportion

Confidence intervals - Background

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Confidence intervals for a proportion

The sample proportion $\hat{p}$ provides a single plausible value for the population proportion **p**. However, the sample proportion isn't perfect and will have some standard error associated with it. When stating an estimate for the population proportion, it is better practice to provide a plausible range of values instead of supplying just the point estimate.

If we report a point estimate $\hat{p}$, we probably will not hit the exact population proportion. On the other hand, if we report a range of plausible values, representing a confidence interval, we have a good shot at capturing the parameter.

Constructing a 95% confidence interval

Our **sample proportion** $\hat{p}$ is the most plausible value of the population proportion, so it makes sense to build a confidence interval around this point estimate. The **standard error** provides a guide for how large we should make the confidence interval.

The standard error represents the standard deviation of the point estimate, and when the **Central Limit Theorem** conditions are satisfied, the point estimate closely follows a **normal distribution**.

In a normal distribution, **95%** of the data is within **1.96 standard deviations** of the mean. Using this principle, we can construct a confidence interval that extends 1.96 standard errors from the sample proportion to be 95% confident that the interval captures the population proportion:

Constructing a 95% confidence interval

$$\text{point estimate} \pm 1.96 \times SE$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

What does "95% confident" mean? Suppose we took many samples and built a 95% confidence interval from each. Then about 95% of those intervals would contain the parameter, p. The following figure shows the process of creating 25 intervals from 25 samples from the earlier simulation, where 24 of the resulting confidence intervals contain the simulation's population proportion of p = 0.88, and one interval does not.

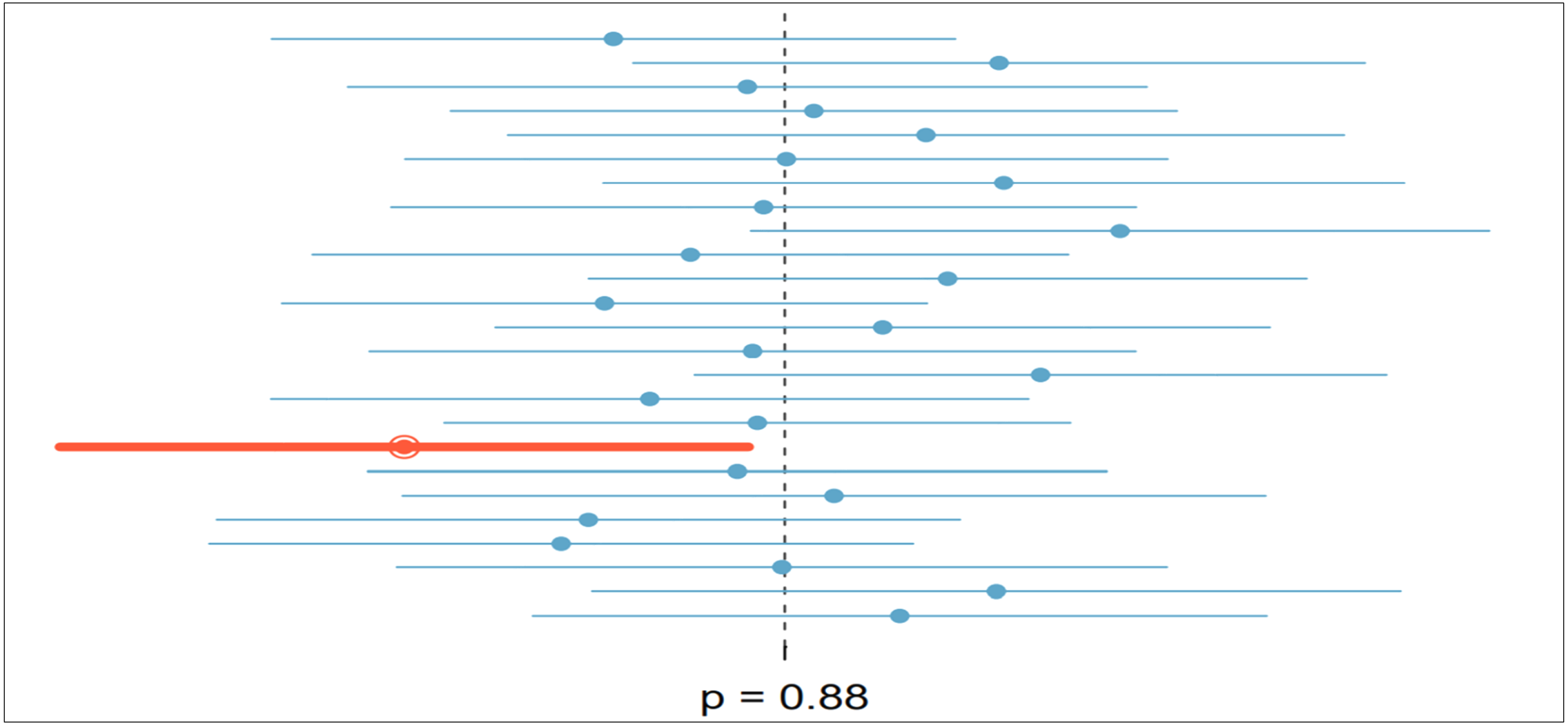


Fig. Twenty-five point estimates and confidence intervals from the simulations in the earlier section. These intervals are shown relative to the population proportion p = 0.88. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been displayed bold.

Example 1

Earlier, we learned about a Pew Research poll where 88.7% of a random sample of 1000 American adults supported expanding the role of solar power. Compute and interpret a 95% confidence interval for the population proportion.

We earlier confirmed that $\hat{p}$ follows a normal distribution and has a standard error of $SE_{\hat{p}} = 0.010$.

To compute the 95% confidence interval, plug the point estimate $\hat{p} = 0.887$ and standard error into the 95% confidence interval formula:

$$\hat{p} \pm 1.96 \times SE_{\hat{p}} \quad \rightarrow \quad 0.887 \pm 1.96 \times 0.010 \quad \rightarrow \quad (0.8674, 0.9066)$$

We are 95% confident that the actual proportion of American adults who support expanding solar power is between 86.7% and 90.7%.

Example 2: Facebook's categorization of user interests

Most commercial websites (e.g. social media platforms, news out- lets, online retailers) collect a data about their users' behaviors and use these data to deliver targeted content, recommendations, and ads. To understand whether Americans think their lives line up with how the algorithm-driven classification systems categorizes them, Pew Research asked a representative sample of **850** American Facebook users how accurately they feel the list of categories Facebook has listed for them on the page of their supposed interests actually represents them and their interests. **67%** of the respondents said that the listed categories were accurate. Estimate the true proportion of American Facebook users who think the Facebook categorizes their interests accurately.

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Example 2: Facebook's categorization of user interests

$$\hat{p} = 0.67 \qquad n = 850$$

The approximate **95% confidence interval** is defined as

$$point\ estimate \pm 1.96 \times SE$$

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.67 \times 0.33}{850}} \approx 0.016$$

$$\begin{aligned} \hat{p} \pm 1.96 \times SE &= 0.67 \pm 1.96 \times 0.016 \\ &= (0.67 - 0.03, 0.67 + 0.03) \\ &= (0.64, 0.70) \end{aligned}$$

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation

point estimate ± 1.96 × SE

Then about **95%** of those intervals would contain the true population proportion (***p***).

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

If the interval is too wide it may not be very informative.

Changing the confidence level

$$\text{point estimate} \pm z^\star \times SE$$

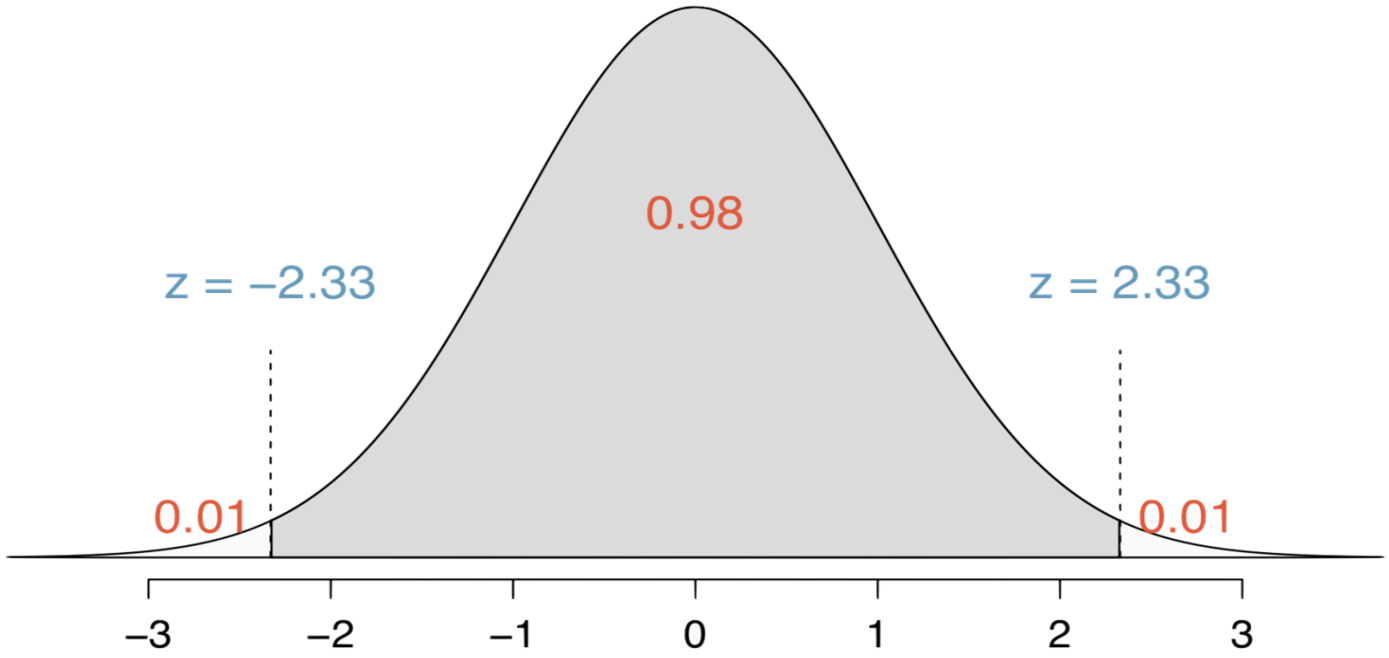
- In a confidence interval, $z^\star \times SE$ is called the **margin of error**, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(a) $Z = 2.05$
(b) $Z = 1.96$
(c) $Z = 2.33$
(d) $Z = -2.33$
(e) $Z = -1.65$

Which of the below Z scores is the appropriate z⋆ when calculating a 98% confidence interval?

(a) Z = 2.05
(b) Z = 1.96
(c) Z = 2.33
(d) Z = −2.33
(e) Z = −1.65



Changing the confidence level

Let us consider confidence intervals where the confidence level is higher than 95%, such as a confidence level of 99%. To create a 99% confidence level, we must widen our 95% interval. On the other hand, if we want an interval with lower confidence, such as 90%, we could use a slightly narrower interval than our original 95% interval.

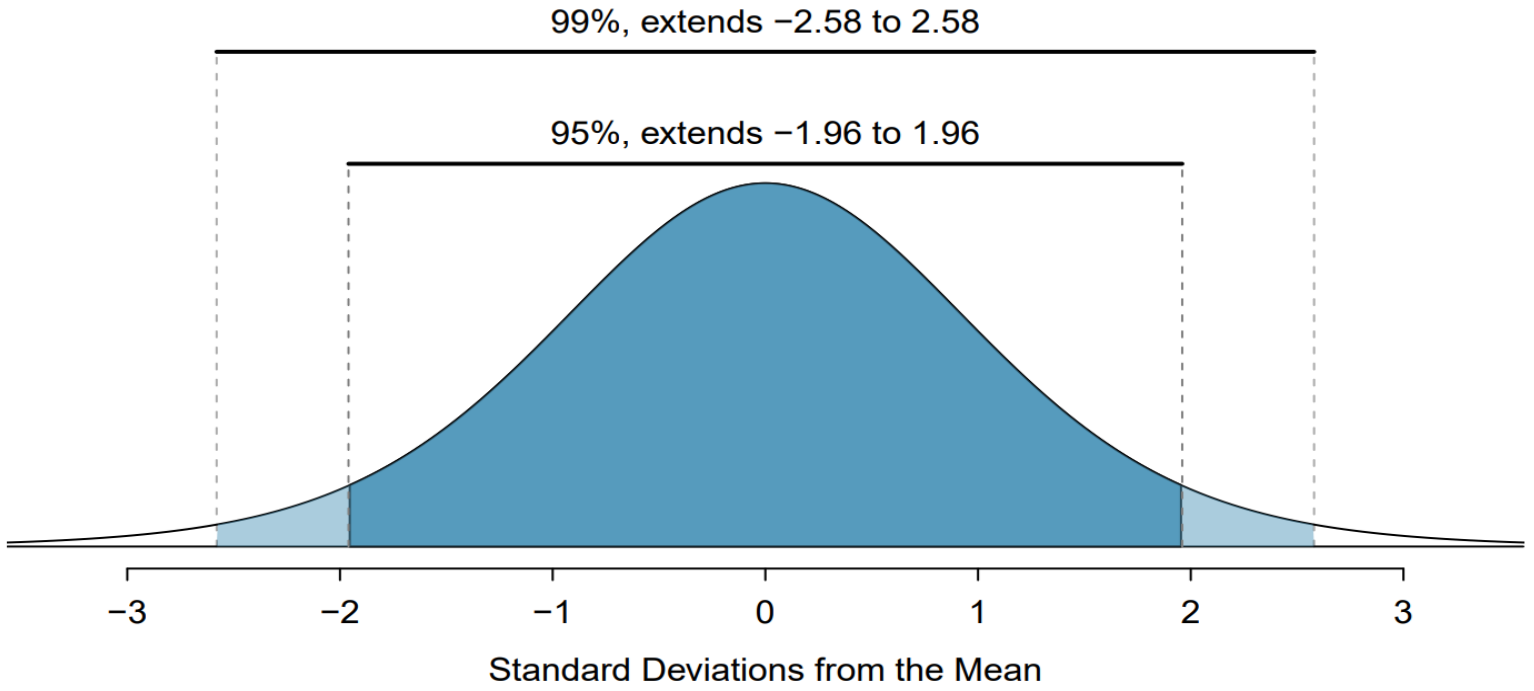
The 95% confidence interval structure provides guidance in how to make intervals with different confidence levels. The general 95% confidence interval for a point estimate that follows a normal distribution is

$$\textbf{point estimate} \pm 1.96 \times \textbf{SE}$$

Changing the confidence level

There are three components to the interval: the point estimate, "1.96", and the standard error. The choice of 1.96 × SE was based on capturing 95% of the data since the estimate is within 1.96 standard errors of the parameter about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level. To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58.

The formula for a 99% confidence interval is **point estimate ± 2.58 × SE**



Interpreting confidence intervals

Confidence intervals are ...

- ***always about the population parameter***
- are not probability statements
- only about population parameters, **not individual observations**
- only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter

Hypothesis Testing for a Proportion

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

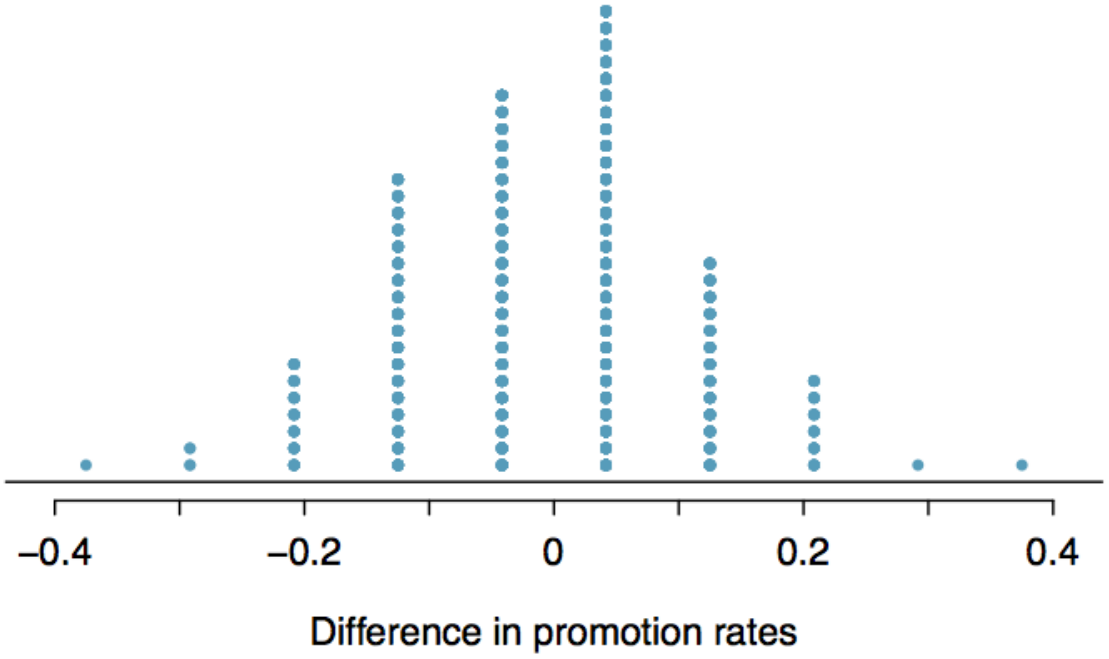
$\hat{p}_{males} = 21 / 24 = 0.88$

$\hat{p}_{females} = 14 / 24 = 0.58$

Possible explanations:

- Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance.
  → null (nothing is going on)
- Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance.
  → alternative (something is going on)

Result



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being 30% or more higher than female promotions), we decided to reject the null hypothesis in favor of the alternative.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

Example 1

A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: ***even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true***. Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

Example 1 – contd.

The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

Example 2

An example of a real-world hypothesis testing situation is evaluating whether a new drug is better or worse than an existing drug at treating a particular disease. What should we use for the null and alternative hypotheses in this case?

The null hypothesis (H_0) in this case is the declaration of no difference: the drugs are equally effective.

The alternative hypothesis (H_A) is that the new drug performs differently than the original, i.e. it could perform better or worse.

Testing hypotheses using confidence intervals

A 95% confidence interval for the proporton of American Facebook users who think Facebook categorizes their interests accurately as 64% to 67%. Based on this confidence interval, do the data support the hypothesis that **majority** of American Facebook users think Facebook categorizes their interests accurately.

The associated hypotheses are:

H_0: $p = 0.50$: 50% of American Facebook users think Facebook categorizes their interests accurately

H_A: $p > 0.50$: More than 50% of American Facebook users think Facebook categorizes their interests accurately

Null value is not included in the interval → reject the null hypothesis.

This is a quick-and-dirty approach for hypothesis testing, but it doesn't tell us the likelihood of certain outcomes under the null hypothesis (p-value)

Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors (contd.)

There are two competing hypotheses: the null and the alternative.
In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty - ***Type 2 error***
- Declaring the defendant guilty when they are actually innocent - ***Type 1 error***

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"
- William Blackstone

Significance Level

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject (H_0) unless we have strong evidence. But what precisely does strong evidence mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject (H_0) more than 5% of the time. This corresponds to a significance level of 0.05. That is, if the null hypothesis is true, the significance level indicates how often the data lead us to incorrectly reject (H_0). We often write the significance level using α (the Greek letter alpha): α = 0.05.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $a = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = a$$

- This is why we prefer small values of a -- increasing a increases the Type 1 error rate.

Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favour of the alternative hypothesis. Statistical hypothesis testing typically uses the p-value method rather than making a decision based on confidence

P-VALUE

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, in this section the sample proportion, to help compute the p-value and evaluate the hypotheses.

COMPARE THE P-VALUE TO α TO EVALUATE H_0

When the p-value is less than the significance level, α, reject H_0. We would report a conclusion that the data provide strong evidence supporting the alternative hypothesis.

When the p-value is greater than α, do not reject H_0, and report that we do not have sufficient evidence to reject the null hypothesis.

In either case, it is important to describe the conclusion in the context of the data.

References

1) Ani Adhikari. John DeNero, Computational and Inferential Thinking: The Foundations of Data Science. GitBook, 2019.

2) OpenIntro Statistics online book: OpenIntro Statistics