Packet 8: Sampling Distribution of the Sample Mean Textbook pages: 24 – 30; 701 – 704

After completing this material, you should be able to:

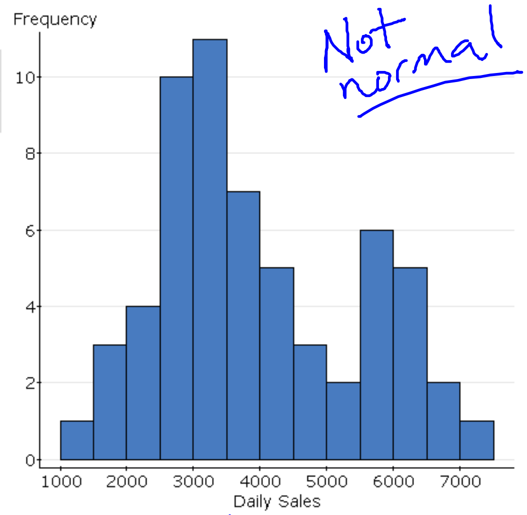
- explain what the symbols $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ represent.
- describe the sampling distribution of the sample mean by discussing its shape, mean, and standard deviation.
- find probabilities associated with various sample means based on the sampling distribution.
- make inferences from the probability and explain the reasoning.

Back to normals.

A local pub is interested in learning more about their daily sales. In order to do this, daily sales (measured in dollars) from the past two months are collected. A histogram and summary statistics are given below. Use these to fill in the description of the distribution below – include units with the values where appropriate.

Summary statistics:

Column	n	Mean	Std. dev.	Median	Min	Max	Q1	Q3
Daily Sales	60	3917.4433	1512.3071	3611.45	1210.5	7068.2	2853.25	5318.05



Not normal

The distribution of Daily sales in dollars has a shape that is bi-modal. The average sales for mean

(n) these 60 days is $3917.44, and 50% of the time median

daily sales are $3611.45 or less. The smallest sales amount min

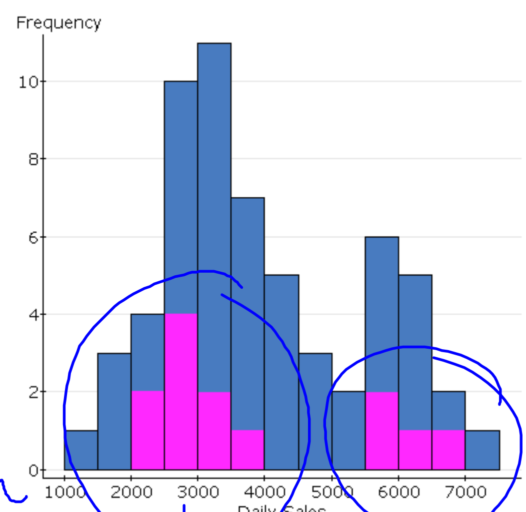
was $1210.50, while the largest max was $7068.20.

IQR, Q1+Q3

The middle 50% of daily sales fell between $2853.25 and $5318.05. According to Chebyshev's Rule, we expect daily sales to fall between $0 and $8454.35 (three standard deviations from the mean) 90 % of the time.

Let's assume this is our population of sales values – what happens if we start to take samples of size 14 from this population of values?

Below is the sample which was taken – the values selected in the sample have been highlighted in pink on the original histogram.



Sample Sales:

2062	2142	2503	2566	2814	2827	3189
3446	3748	4908	5788	5816	6105	6799

What is the mean for this sample of 14 days?

$\bar{y} = \$3908.07$

If we took another sample of 14 days would we expect the same mean? Of course not! It's a random samples; other choices would produce a different mean.

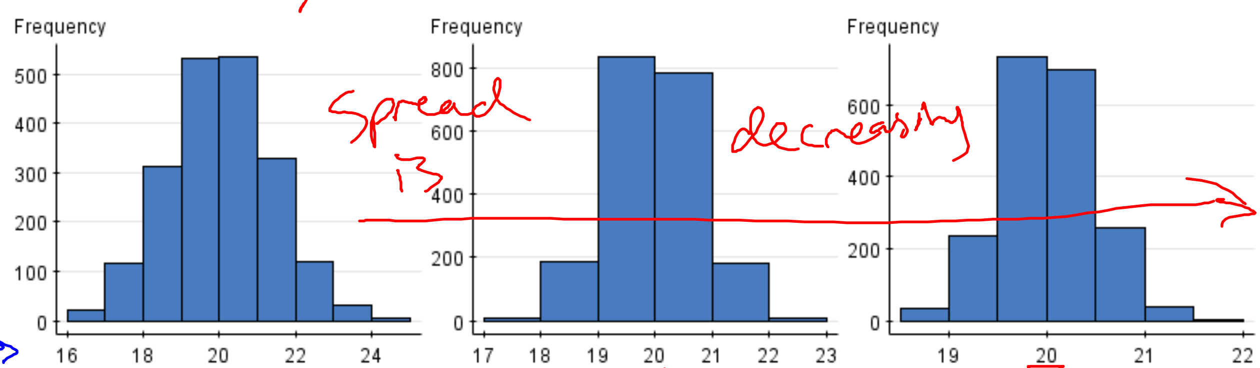
Recall: What is a *sampling distribution* and why is it important? The sampling distribution is the distribution of a statistic obtained by sampling — in this case, $\bar{y}$; the sampling distribution will be more normal than the dist. of y.

We need to understand how the samples vary. To do this, we need to **describe** the sampling distribution of the sample mean. To do this, the following three characteristics must be addressed:

- shape — We like normality. If $n \geq 30$, we will assume that the sampling dist. is normal.
- center — $\mu_{\bar{y}} = \mu_y = \mu$ μ - mean of y
- spread — $\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}}$ $\sigma_y = \sigma$ - std. dev. of y; n - sample size

Example: Each of the following histograms represents a sampling distribution of the sample mean for various sample sizes. Samples of size 5, 15, and 40 were taken from some population. Match each histogram to the appropriate sample size.

μ must be around 20 ...



y-axis is changing

spread is decreasing

sample size: 5 sample size: 15 sample size: 40

which tells us which is which.

Based on the sampling distributions shown, what is the (approximate) population mean? **Explain**.

$\mu \simeq 20$ – all distributions are centered on 20 (or so).

The population from which these samples were taken must have had what shape? **Explain**.

These all look normal, even the sample of size 5. So the distribution of y must have been pretty normal to begin with!

$\mu = 27.2$

Example: It was reported that the average age for a Major League Baseball player is 27.2 years. With the end of the 2016 season, a sports enthusiast wants to know if the average age for players on the Cincinnati Reds is lower. To test this claim, a sample of 16 players on the Reds' roster will be taken. Assume the ages of all MLB players are normally distributed with a standard deviation of 3.12 years.

$n = 16$ $\sigma = 3.12$

y is normally distributed, so $\bar{y}$ is normally distributed as well.

What variable will be recorded in the sample selected? Is this quantitative or categorical?

y-values – individual ages

Only we're collecting 16 to make $\bar{y}$.

Assign the appropriate notation to the values reported in the example.

$\mu = 27.2$ years
$\sigma = 3.12$ years
$n = 16$

What conjecture has been made by the sports enthusiast?

The mean age of the sample obtained from Red's roster is less than 27.2 years : $\mu_{\bar{y}} < \mu$.

Describe the sampling distribution of the sample mean age which will be obtained from a sample of 16 players.

- shape – normal, since y is normally distributed.
- center : $\mu_{\bar{y}} = \mu = 27.2$ years
- spread : $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{3.12}{\sqrt{16}} = \dfrac{3.12}{4} = .78$ years

The fact that the sampling distribution of the sample mean is *normally distributed* is important – we know how to find probabilities from the normal distribution. To do this, though, we will need to modify our formula for the z-score to reflect that we are now dealing with sample means:

Formula Alert!!
This formula will be given on the formula sheet.

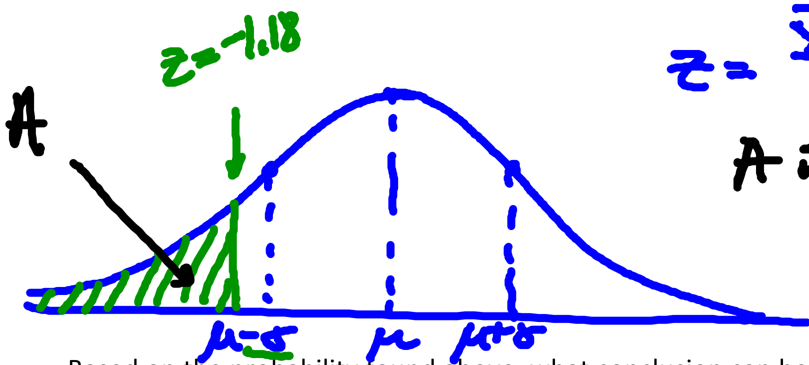
$$z = \frac{\text{obs} - \text{mean}}{\text{std. dev.}} = \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}}$$

$$\boxed{z = \frac{\bar{y} - \mu}{(\sigma/\sqrt{n})}}$$

Let's go back to the example and see how this formula is used …

Back to the example: The sample of 16 players was taken and the average age was found to be 26.28 years. What is the probability of observing this sample mean or some smaller value?

$\bar{y}$



$$z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} = \frac{26.28 - 27.2}{.78} = -1.18$$

A is the probability of a value of $z = -1.18$ (or lower – more extreme) $= .1190$

Based on the probability found above, what conclusion can be drawn?

While the result is to the left ($\bar{y} < \mu$, so z is negative), it is not extremely left: a value of z this extreme or more so occurs .1190 of the time. So we would fail to reject a null of equality of $\mu_{\bar{y}}$ & μ at any typical α level.

$\mu = 10.19$

Example: According to Nielsen, the 2010 NCAA Men's Basketball Tournament (aka March Madness) averaged 10.19 million viewers (including online viewers). The NCAA has expanded its online coverage in recent years, and it is thought the mean number of viewers will have increased. Assume the population is right-skewed with a standard deviation of 6.15 million viewers.

$\sigma = 6.15$ not normal.

What conjecture has been made by the NCAA?

That the mean μ will have increased due to their expansive & exceptional coverage.

A sample of 40 tournament games is taken (from 2011 – 2014), completely describe the sampling distribution of the sample mean number of viewers? $\rightarrow \bar{y}$

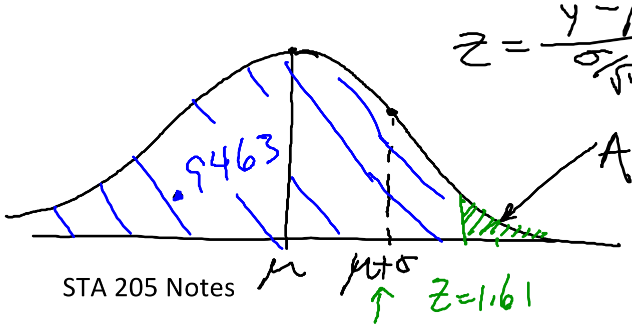
$n = 40$

- shape: normal ($n \geq 30$) This heals almost any wonky distribution of y (e.g. right-skew)
- center: $\mu_{\bar{y}} = \mu = 10.19$
- spread: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{6.15}{\sqrt{40}} = .9724$ million viewers

The sample of 40 games has a mean of 11.76 million viewers. What is the probability of observing an average of 11.76 or larger?

$\bar{y} = 11.76$ (good news! $\bar{y} > \mu$ – is it enough greater?)

$$z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} = \frac{11.76 - 10.19}{.9724} = 1.61$$



A – prob. of a value of z this great or greater $= 1 - .9463$ $= \boxed{.0537}$

Based on the probability computed above, what can be concluded about the mean number of viewers for March Madness?

A probability of .0537 is fairly extreme – but with a typical α of .05 we would not be able to reject a null of equal means ($\mu_{\bar{y}} = \mu = 10.19$).

If the sample size were increased to 75 games, between what two values would we expect **68%** of the sample means to fall?

$\bar{y}$ should be normally distributed. Tip-off that we could use the empirical rule.

We would expect 68% of values to fall between $\mu_{\bar{y}} - \sigma_{\bar{y}}, \mu_{\bar{y}} + \sigma_{\bar{y}}$: $10.19 \pm \frac{6.15}{\sqrt{75}}$

$10.19 \pm .71$: $[9.48, 10.90]$

Example: An ambulance service reports that the mean time required for ambulances to reach their destinations was 10 minutes with a standard deviation of 4.72 minutes. Given traffic congestion it has been hypothesized that the mean time has increased.

$\mu = 10$ $\sigma = 4.72$ No info about dist. of y.

If a sample of 35 ambulance runs is selected, how would the sampling distribution of the sample mean time to destination be described? $n = 35$ $n \geq 30$ heals all wonkiness!

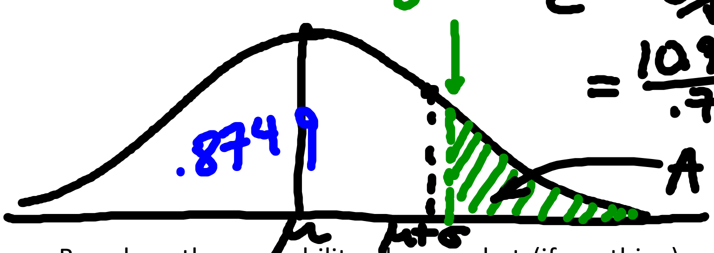
– shape: with $n \geq 30$, expect $\bar{y}$ to be distributed normally.

– center: $\mu_{\bar{y}} = \mu = 10$

– spread: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{4.72}{\sqrt{35}} = .798$ minutes

Suppose that a sample of 35 ambulance runs is taken, and they report a sample mean time to the destinations of 10.92 minutes. What is the probability of observing a sample mean time of 10.92 minutes or more?

$\bar{y} = 10.92$ (it's bigger than 10 – significantly so? surprisingly so?)



$$z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} = \frac{10.92 - 10}{.798} = \frac{.92}{.798} = 1.15$$

A – prob. of a z that extreme or more so:

$= 1 - .8749 = .1251$

Based on the probability above, what (if anything) can be inferred about the mean time required for ambulances to reach their destination?

While $\bar{y}$ leans higher, it's not dramatically higher: we would get a result this high or higher 12.5 times out of a hundred. I wouldn't feel comfortable concluding that mean time has increased.