# Stat 88: Prob. & Mathematical Statistics in Data Science

Prob dsn of $X_k$

$X_1, X_2, \ldots, X_n$
iid r.v.

Distribution of $\frac{S_n}{n}$

$\overline{X_n}, A_n$
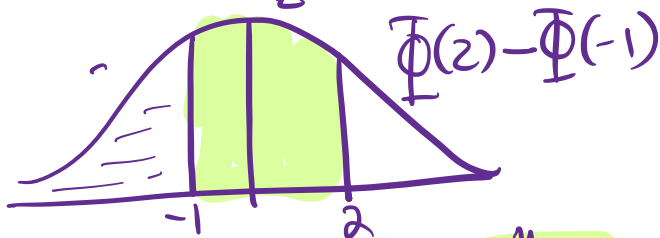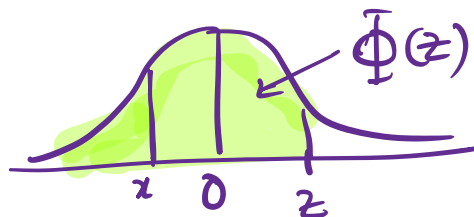


Lecture 20 PART 1: 4/5/2022

Finishing up chapter 8 and the Central Limit Theorem

# Standard normal cdf

$\Phi$

- $\Phi(z) = \int_{-\infty}^{z} \phi(x)dx$, where $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, $-\infty < x < \infty$

$\Phi(x) \hookleftarrow F(x)$
cdf

$\phi(x) \hookleftarrow f(x)$
prob. density function

$\Phi(z)$ is the area under the curve $y = \phi(x)$.

$\phi(x)$

$\Phi(z)$
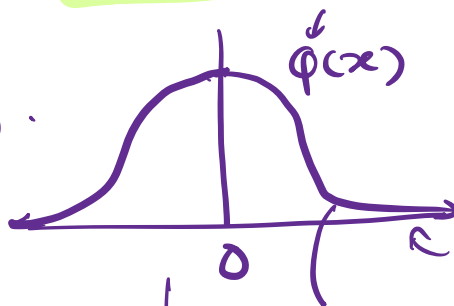
$x$ axis is asymptotic to $\phi(x)$

Standard Normal Distribution $N(0,1)$

$X \sim N(0,1)$

$\Phi(2) - \Phi(-1)$

This means that
$E(X) = 0 \cdot (\mu)$
$Var(X) = 1 \ (\sigma^2)$
$P(X=x) = 0$ for all $x$

$P(X \leq y) = \Phi(y) = \int_{-\infty}^{y} \phi(x)dx$
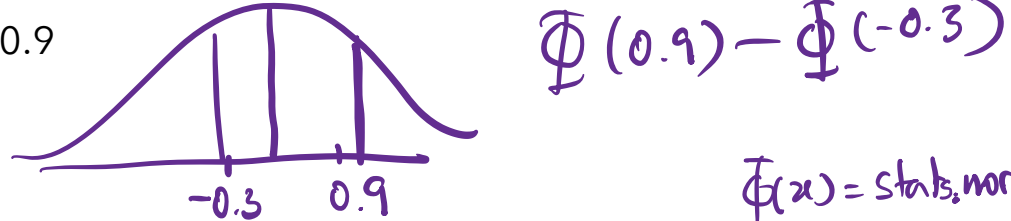
$\Phi(-z) = \Phi(z)$?

4/4/22

2

$-z \quad 0 \quad z \quad \Phi(-z) = 1 - \Phi(z)$

# Standard normal cdf, symmetries, percentiles

Example: Find the area
(a) to the right of 1.25

$= 1 - \Phi(1.25)$

(b) between -0.3 and 0.9

$\Phi(0.9) - \Phi(-0.3)$

$\Phi(x) = $ stats.norm.cdf($x$)
$\qquad$ pnorm($z$)

(c) outside -1.5 and 1.5.

$2 \times \Phi(-1.5) =$

$= \Phi(-1.5) + (1 - \Phi(1.5))$

$-0.95$
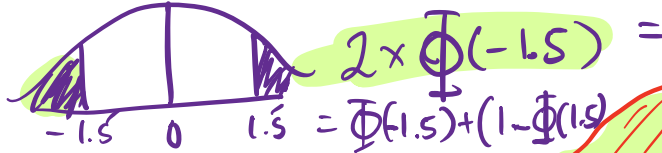
2. (a) Find $z$ such that $\Phi(z) = 0.95$

qnorm($x$)
stats.norm.ppf($x$)

$\Phi^{-1}(0.95)$

$z = ??$
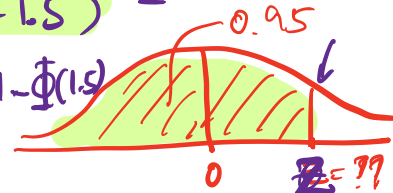
$\uparrow$ 95th percentile
aka 0.95th quantile.

(b) Find $z$ so that the area in the **middle** is 0.95 $\boxed{\Phi(z) - \Phi(-z) = 0.95}$

$\Phi, \Phi^{-1}$

0.025

$-z \quad 0 \quad z$

$\approx 1.96$

$\Phi^{-1}(0.025) = -1.96$
$\Phi(0.975) = 1.96$

# The Central Limit Theorem

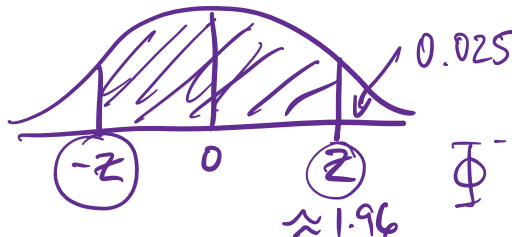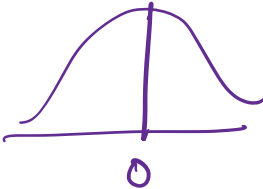- Suppose that $X_1, X_2, \ldots, X_n$ are iid with mean $\mu$ and SD $\sigma$

- Let $S_n = X_1 + \cdots + X_n$ be the sample sum, and $A_n = \frac{S_n}{n}$ be the sample mean

- Then the distribution of $S_n$ (and $A_n$) is **approximately normal** for large enough $n$.

- For $S_n$, the distribution is approximately normal (bell-shaped), centered at $E(S_n) = n\mu$ and with spread given by by $SD(S_n) = \sqrt{n}\,\sigma$.

- For $A_n$, the distribution is centered at $E(A_n) = \mu$ with spread $SD(A_n) = \sigma/\sqrt{n}$

$$\begin{cases} \dfrac{S_n - n\mu}{\sqrt{n}\,\sigma} & \sim N(0,1) \\[2ex] \dfrac{A_n - \mu}{\sigma/\sqrt{n}} & \sim N(0,1) \end{cases}$$

# Normal approximations : standard units

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Let $X$ be any random variable, with expectation $\mu$ and SD $\sigma$, consider a new random variable that is a linear function of $X$, created by shifting $X$ to be centered at 0, and dividing by the SD. If we call this new rv $X^*$, then $X^*$ has expectation ___0___ and SD ___1___.

- $E(X^*) =$ 0
- $SD(X^*) =$ 1

  $\Big\}$ Show this by using $X^* = \dfrac{X - \mu}{\sigma}$ & rules for expectation & SD

- This new rv does not have units since it measures how far above or below the average a value is, in SD's. Now we can compare things that we may not have been able to compare.

- Because we can convert anything to standard units, ***every normal curve is the same.***

# The many normal curves → the *standard normal* curve
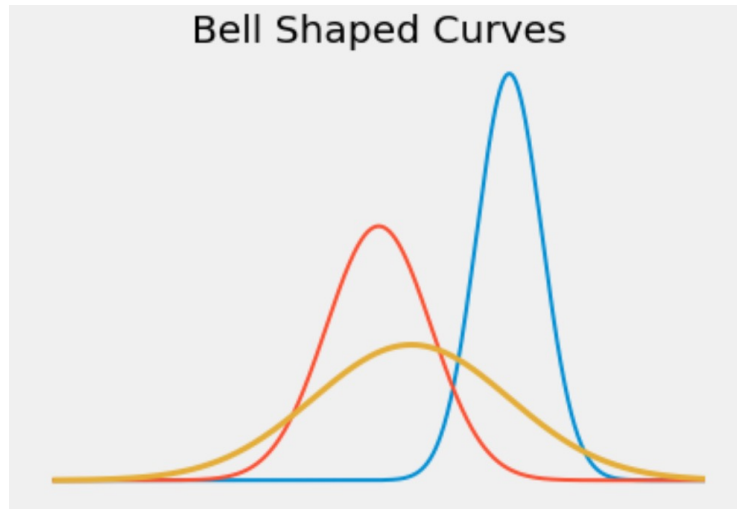
### Bell Shaped Curves

- Just one normal curve, standard normal, centered at 0. All the rest can be derived from this one.

# How do we approximate the area of a range in a histogram?
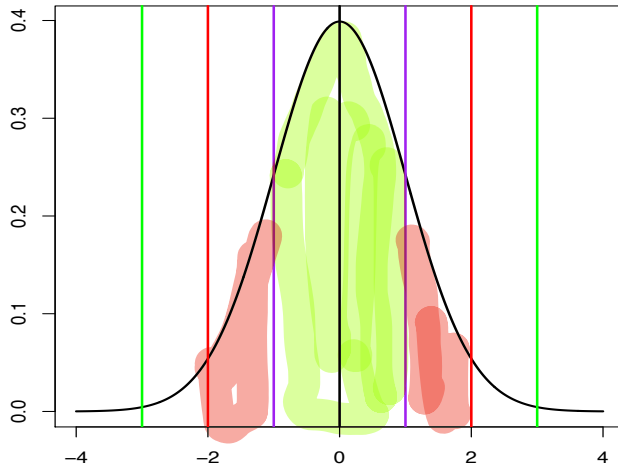


- Many histograms bell-shaped, but not on the same scale, and not centered at 0

- Need to convert a value to **standard units** – see how many SDs it is above or below the average

- Then we can **approximate** the area of the histogram using the area under the standard normal curve (using for example, stats.norm.cdf for the actual numerical computation)

- <mark>68%-95%-99.7% rule</mark>

  (**Empirical rule**)

Total area under the curve = 100%
Curve is symmetric about 0
The areas between 1, 2 and 3 SDs away:
Between -1 and 1 the area is 68.27%
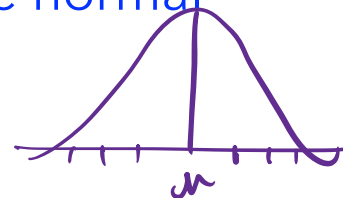Between -2 and 2 the area is 95.45%
Between -3 and 3 the area is 99.73%

exercise : verify using R/ Python/ online calc.

# How to decide if a distribution could be normal

- Need enough SDs on both sides of the mean.

- In 2005, the murder rates (per 100,000 residents) for 50 states and D.C. had a mean of 8.7 and an SD of 10.7. Do these data follow a normal curve?

  NO!

- If you have indicators, then you are approximating binomial probabilities. In this case, if $n$ is very large, but $p$ is small, so that $np$ is close to 0, then you can't have many sds on the left of the mean. So need to increase $n$, stretching out the distribution and the n the normal curve begins to appear.

- If you are not dealing with indicators, then might bootstrap the distribution of the sample mean and see if it looks approximately normal.

# Stat 88: Prob. & Mathematical Statistics in Data Science



https://xkcd.com/892/

Lecture 20 PART 2: 4/5/2022

Section 9.1, 9.2

Hypothesis tests & A/B testing

# Hypothesis tests

- Hypothesis tests or *tests of significance* are tests in which we use data to draw conclusions, or *make inferences* about the process that generated the data, or the population from which we drew the sample.

- Underlying idea: Observed values can't be too far from the expected value, *if our assumptions are correct.*

- What if they are not correct? *How far is too far*?

- Toss a coin 100 times. See 54 heads. Do you have reason to believe that the coin is not fair?

- What about 60 heads? 66 heads?

- How would we decide?

Assumption coin is fair

$X = \# $ of $H$, $X \sim Bin(100, 0.5)$

in 100     $E(X) = 50$

$SD(X) = 5$

- Let's look at a couple of examples, and review the ideas of hypothesis testing

# Example: Gender discrimination?

A large supermarket chain occasionally selects employees to receive management training. A group of women there claimed that female employees were passed over for this training in favor of their male colleagues. The company denied this claim.

Suppose that the large employee pool of the Florida chain that can be tapped for management training is half male and half female.

*Let's assume 1000 employees, 500M, 500F*

Since this program began, *none* of the 10 employees chosen have been female. What would be the probability of 0 out of 10 selections being female, if there truly was no gender bias?

Let $X = \#$ of Women in 10 chosen

$X \sim HG(N=1000, G=500, n=10)$

OR $X \sim Bin(n=10, P=\frac{1}{2})$

$P(X=0)$ very small $\approx 0.0009$.

## Example: Woburn

*Intro to Stats by DeVeaux, Velleman & Bock*

In the early 1990s, a leukemia cluster was identified in the Massachusetts town of Woburn. Many more cases of leukemia, a malignant cancer that originates in a bone marrow cell, appeared in this small town than would be predicted. Was this evidence of a problem in the town or just chance?

(Let's say there were 8 cases that year.)   (A Civil Action)

We make an assumption: the high rates of Leukemia in Woburn are just by chance

$X = \#$ of people in Woburn suffering from Leukemia

$X \sim Bin(n, p)$   $n =$ Woburn popn

$p =$ Leukemia rate in US at that time

(Assumption: each resident is indep of every other resident & has same chance of leukemia)

$n = 35000$,   $p = \dfrac{30\,800}{280 \times 10^6} = 0.00011$   $\left(\dfrac{\#\text{ of cases in that year}}{\text{US population}}\right)$

4/4/22                                                                                     12

Exercise: Complete this problem by computing $P(X \geq 8)$

# Hypotheses tests: Review of steps

1. State the *null* hypothesis ($H_0$) – that is, what is the assumption we are going to make. This will determine how we compute probabilities

2. State an *alternative* hypothesis ($H_A$) Note that this should not overlap with the null hypothesis, and it may or may not define probabilities (example: there was gender bias etc)

3. Decide on a test statistic to use that will help you decide which of the two hypotheses is supported by the evidence (data). Usually there is a natural choice. Use the null hypothesis to specify probabilities for the test statistic.

4. Find the observed value of the test statistic, and see if it is *consistent* with the null hypothesis. That is, compute the chance that we would see such an observed value, or more extreme values of the statistic (*p-value*).

5. State your conclusion: whether you reject the null hypothesis or not. This is based on your chosen cutoff ("level of the test"). Reject if the $p$-value is less than the cutoff.

# Observed significance levels (a.k.a *p*-values)

*(handwritten annotations: dsn for test statistic, p-value, obs value)*
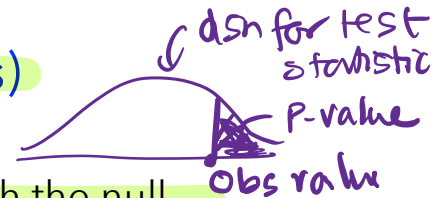
- The *p-value* decides if observed values are *consistent* with the null hypothesis. It is a *tail* probability (also called *observed significance level)*, and is the chance, **assuming that the null hypothesis is true**, of getting a test statistic equal to the one that was observed or even more in the direction of the alternative.

- If this probability is too small, then something is wrong, perhaps with your assumption (null hypothesis). That is, the data are unlikely if the null is true and therefore, your data are ***inconsistent*** with the null hypothesis.

- *p*-value is **not** the chance of null being true. The null is either true or not.

- The *p*-value is a *conditional probability* since it is computed *assuming* that the null hypothesis is true.

- The smaller the *p*-value, the stronger the evidence ***against*** the null and ***towards*** the alternative (in the direction of the alternative).

-  Traditionally, below 5% ("result is statistically significant") and 1% ("result is highly significant") are what have been used. Significant means the *p*-value is small, not that the result is important.