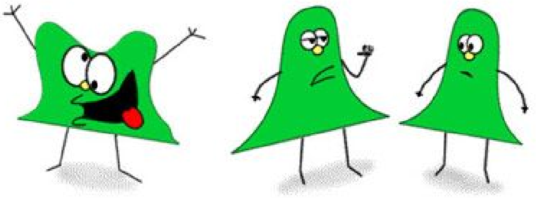
CSCI 3022-002 Intro to Data Science
The Central Limit Theorem

What were the Python commands for the standard normal distribution?

Announcements and Reminders

- Practicum due Monday!



"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW...NORMAL!"

The Normal Distribution

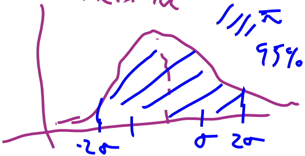
1. **Definition:** *Normal Distribution:*

   A continuous r.v. X is said to have a *normal distribution* with parameters μ and $\sigma^2 > 0$, if the pdf of X is:

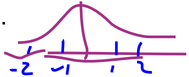
   Variance

   mean μ

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



2. The normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$ is called the *standard normal distribution*, and is denoted by Z.



3. The cdf of Z is given by

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

   add up from $-\infty$ to z

4. To access the density function, we use STATS.NORM.PDF. To access the cdf, we use STATS.NORM.CDF.

   → in: x-value, out: probability

   → in: x-value, out: height of fn.

Standard Quantiles

what if we want

in: probabilities
out: x-values

The 99th *percentile* of the standard normal distribution is that value of z such that the area under the z curve to the left of the value is 0.99.

Tables and cdf functions give, for fixed z, the area under the standard normal curve to the left of z; now we have the area and want the value of z.

Prob. that a random normal is less than z

This is the "inverse" problem to $\boxed{F(z)} = P(Z \leq z) = \boxed{\Phi(z)}$. Now we're asking: what is the x value so that $\Phi(x) = P(Z \leq x) =$ some given probability. We can even write it that way: if the given probability is p, we'd have $x = \Phi^{-1}(p)$.

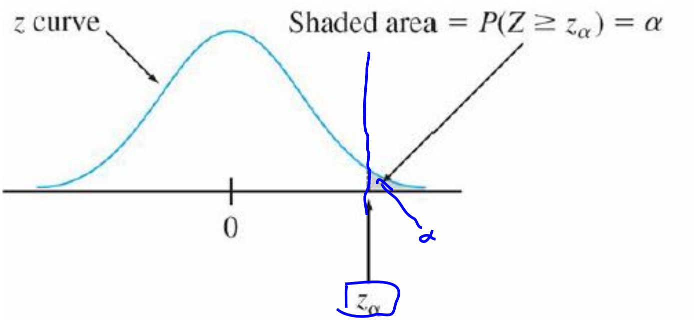
$\Phi(x) = p$ $\quad \Phi(x) = .99 \Longleftarrow \quad x = \Phi^{-1}(.99)$

To access these, we use STATS.NORM.PPF. ppf stands for "$\underbrace{\text{percentile}}_{\text{in}}$ $\underbrace{\text{point}}_{\text{out}}$ function," as in it returns the point that is e.g. the 95th percentile of Z!.

Standard Quantiles

In statistical inference, we need the z values that give certain tail areas under the standard normal curve.

There, this notation will be standard: z_α will denote the z value for which α of the area under the z curve lies to the right of z_α.

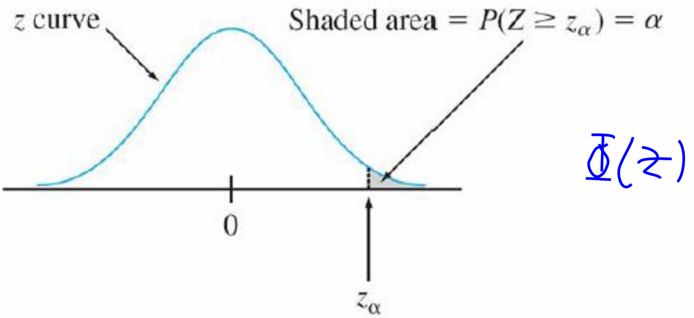


Why this? How does this relate to the *cdf*? → cdfs count area to the LEFT

Standard Quantiles

In statistical inference, we need the z values that give certain tail areas under the standard normal curve.

There, this notation will be standard: $\underline{z_\alpha}$ will denote the z value for which $\underline{\alpha}$ of the area under the z curve lies to the right of $\underline{z_\alpha}$.



$\Phi(z) = P(Z \leq z)$

$= 1 - P(Z > z)$

Why this? How does this relate to the *cdf*? $\boxed{P(Z \geq z_\alpha) = \alpha} \Longrightarrow \Phi(z_\alpha) = 1 - \alpha$

Non-Standard Normals

When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing." The standardized variable is:

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

recall: $X-\mu_x$: mean 0
then dividing by σ
makes s.d. = 1.

Proposition: If X has a normal distribution with mean μ and standard deviation σ, then

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

is distributed standard normal.

Non-Standard Normals

When $\underline{X \sim N(\mu, \sigma^2)}$, probabilities involving X are computed by "standardizing." The standardized variable is:

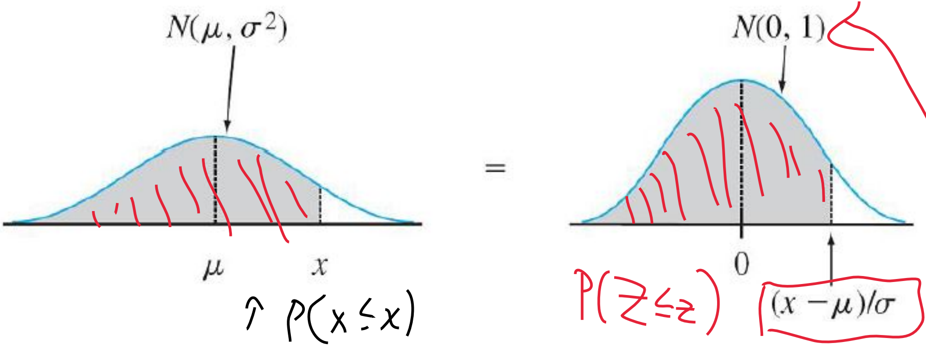
$$Z = \frac{X - \mu}{\sigma}$$

Proposition: If X has a normal distribution with mean $\underline{\mu}$ and standard deviation $\underline{\sigma}$, then

is distributed standard normal.

Non-Standard Normals

Why do we standardize normal random variables?



Equality of nonstandard and standard normal curve areas

lives in our software

Using Normals

Example:
The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.

Research suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

What is the probability that reaction time is between 1.00 sec and 1.75 sec?

Solution:

Example: For a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

What is the probability that reaction time is between 1.00 sec and 1.75 sec?

Easiest ans: norm.cdf: $P(\text{reaction} \le 1s) = \text{norm.cdf}(1, \text{mean}=1.25, \text{scale}=.46)$

$P(\text{reaction} \le 1.75s) = \text{" "}(1.75, \ldots)$

Instead we normalize: GOAL: $\boxed{P(1 \le X \le 1.75)}$

where $X \sim N(1.25, (.46)^2)$

we can only do $P(a \le Z \le b)$

X becomes standard if: $\frac{X-\mu}{\sigma}$ $P(1 \le X \le 1.75) = P\left(\frac{1-1.25}{.46} \le \frac{X-1.25}{.46} \le \frac{1.75-1.25}{.46}\right)$

Solution:

Example: For a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

$$X \sim N(1.25, .46^2)$$

What is the probability that reaction time is between 1.00 sec and 1.75 sec?
We want $P(1 < X < 1.75)$... but we can't compute these probabilities unless the r.v. in the middle of the inequality is *standard* normal. So we normalize!

Solution:

Example: For a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

What is the probability that reaction time is between 1.00 sec and 1.75 sec?
We want $P(1 < X < 1.75)$... but we can't compute these probabilities unless the r.v. in the middle of the inequality is *standard* normal. So we normalize!

$$P(1 < X < 1.75) = P(1 - 1.25 < X - 1.25 < 1.75 - 1.25)$$

$$= P\left(\frac{-.25}{.46} < \frac{X - 1.25}{.46} < \frac{.5}{.46}\right) = P\left(\underbrace{\frac{-.25}{.46}}_{a} < Z < \underbrace{\frac{.5}{.46}}_{b}\right)$$

$$= \Phi\left(\frac{-.25}{.46}\right) + \Phi\left(\frac{.5}{.46}\right)$$

Stats.cdf(.5/.46) − Stats.cdf(−.25/.46)

iid

Definition: *Random Sample:*

The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size n if:

1.

2.

We say that these X_i's are:

iid

Definition: *Random Sample:*
The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size n if:

1. $X_1, X_2, \ldots X_n$ are independent.

2.No value in the population has a higher chance of being included than any other.

We say that these X_i's are: *independent* and *identically distributed.*
and we write:

$$X_1, X_2, \ldots X_n \boxed{\overset{iid}{\sim}} f(x;\theta)$$

(θ is the parameters)

f is the pdf for each & every X

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample.

Examples?

Mean

Sample variance

median

Proportion of time node 0 was "I" ...

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample.

Examples?

1. *Sample* Mean might estimate a population mean.
2. *Sample* Variances estimate population variance.
3. *Sample* Quantiles
4. $\hat{p}$ for p

   population proportion

   true (e.g. Bernoulli) probability
5. etc., etc.

$$\underbrace{\bar{X} = \frac{\sum x_i}{n}}_{\text{data \& practice}} \overset{?}{\approx} \underbrace{\int x f(x)\,dx}_{\text{prob. theory}}$$

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample.

Examples?

1. *Sample* Mean might estimate a population mean.
2. *Sample* Variances estimate population variance.
3. *Sample* Quantiles
4. $\hat{p}$ for p
5. etc., etc.

Why use one estimator over another?

2 things we want:

our data-based guess should:
1) be "close"
2) get better as n increases.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean $\underline{\bar{x}}$ is a random variable (since it is based on a random sample).

This means that $\underline{\bar{X}}$ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean $\underline{\;\bar{X}\;}$ is a random variable (since it is based on a random sample).

This means that $\underline{\;\bar{X}\;}$ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

1. n

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean ___ is a random variable (since it is based on a random sample).

This means that ___ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

1. n
2. population distribution

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Estimators and Their Distributions

We use estimators to summarize our i.i.d. sample. Any estimator, including the sample mean ___ is a random variable (since it is based on a random sample).

This means that ___ has a distribution of it's own, which is referred to as sampling distribution of the sample mean. This sampling distribution depends on:

1. n
2. population distribution
3. method of sampling

Definition: The standard deviation of this distribution is called the *standard error* of the estimator.

Distribution of the Sample Mean

for $X_1, X_2, \ldots, X_n$

we know: $E[X_i]$, $Var[X_i]$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

avg. value of sample mean

recall: $\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$

$$E[\bar{X}] = E\left[\frac{\sum X_i}{n}\right] = \frac{1}{n} \sum_{i=1}^{n} \underbrace{E[X_i]}_{\mu} = \frac{1}{n}(n \cdot \mu)$$

$= \mu$

$$Var[\bar{X}] = Var\left[\frac{\sum X_i}{n}\right] = \frac{1}{n^2} \sum_{i=1}^{n} \underbrace{Var[X_i]}_{\sigma^2} = \frac{1}{n^2}(n\sigma^2)$$

iid

The standard deviation of the sample mean is:

$Var[\bar{X}] = \frac{\sigma^2}{n}$ so $S.Dev.[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

This is also called the standard error of the mean.

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

$$E[\bar{X}] = \mu$$

$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

The standard deviation of the sample mean is:

This is also called the standard error of the mean.

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

$$E[\bar{X}] =$$

$$Var[\bar{X}] =$$

The standard deviation of the sample mean is:

$$s.e.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

This is also called the standard error of the mean.

Distribution of the Sample Mean

What does this mean? Why is it true?

$$E[\bar{X}] =$$

$$Var[\bar{X}] =$$

Also, what do we know about the *distribution* of the sample mean?

Distribution of the Sample Mean

What does this mean? Why is it true?

$$E[\bar{X}] = E\left[\frac{\sum X_i}{n}\right] = \frac{\sum E[X_i]}{n} = \frac{n\mu}{n} = \mu$$

$$Var[\bar{X}] =$$

Also, what do we know about the *distribution* of the sample mean?

Distribution of the Sample Mean

What does this mean? Why is it true?

"Law of Large numbers" → more data is good (particularly good for $\bar{X}$)

$$E[\bar{X}] = E\left[\frac{\sum X_i}{n}\right] = \frac{\sum E[X_i]}{n} = \frac{n\mu}{n} = \mu$$

$$Var[\bar{X}] = Var[\sum X_i/n] = \frac{1}{n^2}\sum Var[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Also, what do we know about the *distribution* of the sample mean?

$\bar{X} \boxed{\to} \mu$ — limit as $n \to \infty$

Distribution of the Sample Mean (Normal Population)

Proposition: ... if you add up $N(a,b) + N(c,d)$

If $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

get: $N(a+c, \ldots)$

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

Distribution of the Sample Mean (Normal Population)

Proposition:

If $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$\bar{X} : \frac{\sum X_i}{n}$ ← add up normals.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

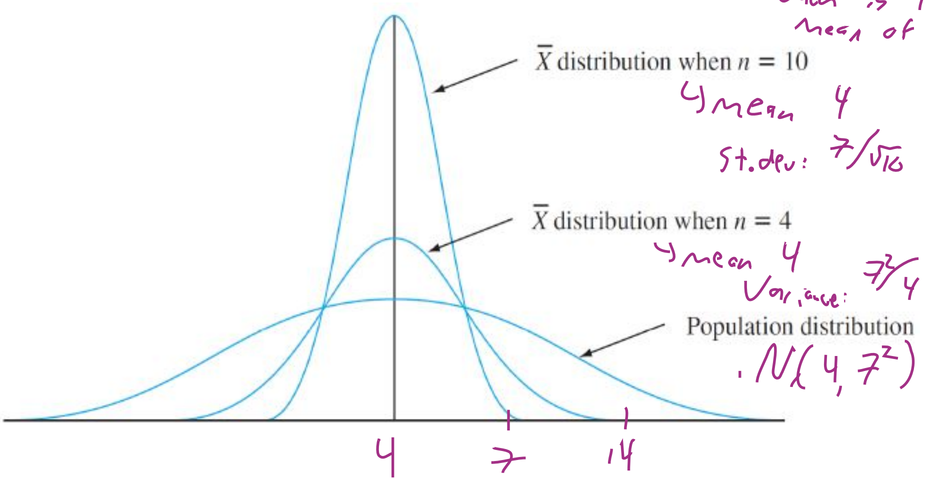
Var $= \sigma^2/n$

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

stats.norm.cdf.

Distribution of the Sample Mean (Normal Population)

...run 100 experiments, each is the mean of "n" normals.



Central Limit Theorem

But what if the underlying distribution of the X_i's is not normal?

Central Limit Theorem

Important: When the population distribution is nonnormal, averaging produces a distribution more bellshaped than the one being sampled.

A reasonable conjecture is that if n is large, a suitable normal curve will approximate the actual distribution of the sample mean.

The formal statement of this result is one of the most important theorems in probability: *Central Limit Theorem!*

Central Limit Theorem

means as $n \to \infty$

Theorem: *Central Limit Theorem:*

Central Limit Theorem

Theorem: *Central Limit Theorem:*
Let $X_1, X_2, \ldots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

Central Limit Theorem

Theorem: *Central Limit Theorem:*
Let $X_1, X_2, \ldots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

mean is "approximately distributed as"

$$N\left(\mu, \underbrace{\sigma^2/n}_{\text{smaller than underlying variance } \sigma^2}\right)$$

same as underlying μ

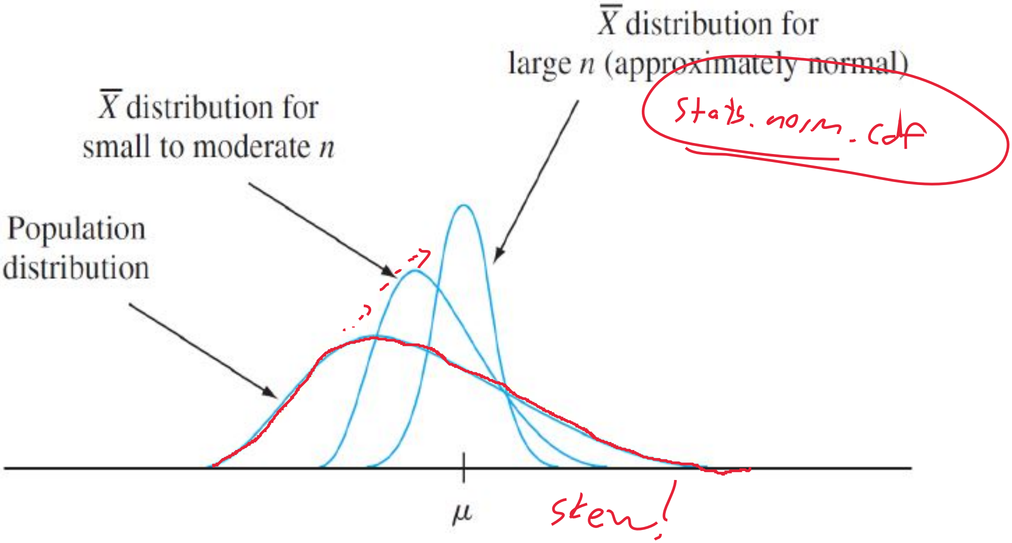
Central Limit Theorem

Theorem: *Central Limit Theorem:*
Let $X_1, X_2, \dots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger the value of n, the better the approximation! Typical rule of thumb: $n > 30$.

Central Limit Theorem



Central Limit Theorem

Example: The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

If 50 batches are independently prepared, what is the (approximate) probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

Example sol:

Example sol:

We want the probability $P(3.5 < \bar{X} < 3.8)$ for $X \sim N(4.0, 1.5)$. Again we normalize... but $\bar{X}$ has much smaller standard deviation than each one of the individual data values!

$$P(3.5 < \bar{X} < 3.8) = P\left(\frac{3.5 - 4.0}{1.5/\sqrt{50}} < \frac{\bar{X} - 4.0}{1.5/\sqrt{50}} < \frac{3.8 - 4.0}{1.5/\sqrt{50}}\right)$$

$$= P\left(\frac{-1}{3/\sqrt{50}} < Z < \frac{-2}{15/\sqrt{50}}\right)$$

for $Z \sim N(0, 1)$ which is

$$\Phi\left(\frac{-2}{15/\sqrt{50}}\right) - \Phi\left(\frac{-1}{3/\sqrt{50}}\right)$$

Central Limit Theorem

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when n is sufficiently large. The problem is that the accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled.

Daily Recap

Today we learned

1. It's all normal? (always has been)

Moving forward:

- nb day Friday

Next time in lecture:

- More: how we can use that it's all normal!