Probability Distributions

Random Variable

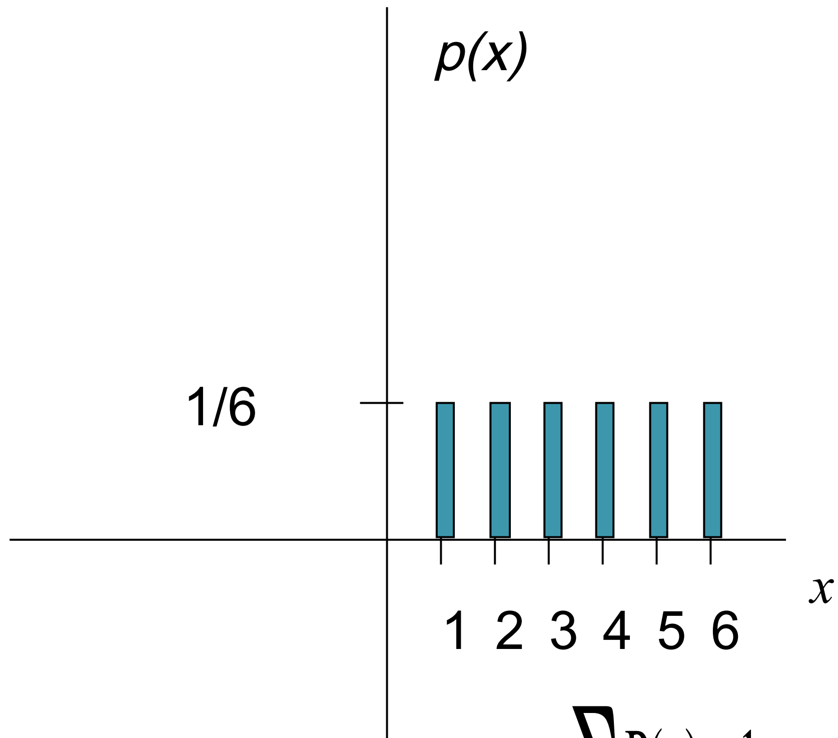
The variable assumes a number of different values, any particular outcome occurs by chance.

- **Discrete** random variables have a countable number of outcomes
- **Continuous** random variables have an infinite continuum of possible values.

For discrete variables, the probability distribution describes the probability of <u>each possible value.</u>

Example: roll of a die

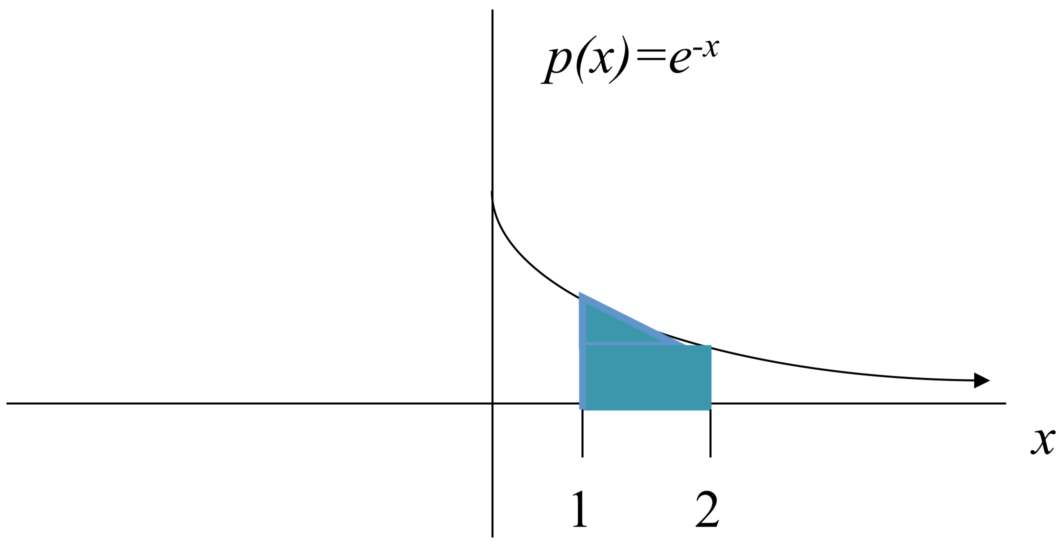
x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$p(x=6)=1/6$



$$\sum_{\text{all } x} P(x) = 1$$

For continuous variables, the distribution describes the probability of a <u>range of values.</u>

Example: Survival after transplant



$$P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23$$

Bernoulli Distribution

$$f(x) = p^{x} (1-p)^{1-x}$$

- Discrete variable
- Two mutually exclusive outcomes:

  success / failure ($x=0,1$)
- One trial

Binomial Distribution

Suppose we repeat a Bernoulli *P* experiment *n* times and count the number *x* of successes, which generalizes to:

$$P(X = x) = \binom{n}{x} p^{x}(1-p)^{n-x}$$

Stata uses k for x

- *p*: probability of "success" in each "trial"
- x: the number of "successes"
- *n*: the number of "trials"

Binomial Distribution

- Assumptions:
  - There are a <u>fixed number</u> of trials *n* <u>(trails are independent)</u>
  - Within each trial, there are <u>two mutually exclusive outcomes</u>
  - The probability of success <u>*p* is constant</u> for each trial

Binomial Distribution

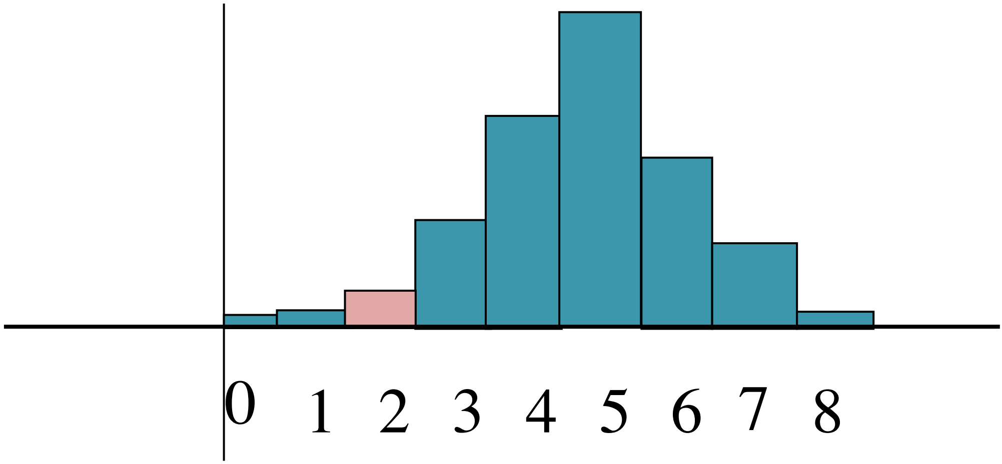
1) $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$

– `di comb(n,x) * p^x * (1-p)^(n-x)`

2) **Stata:** `binomialp(n,k,p)` – P(X=k)

`binomialtail(n,k,p)` – P(X≥k)

Exercise



What is the probability of **2 cases** of disease (p=0.6) in a sample of n=8?

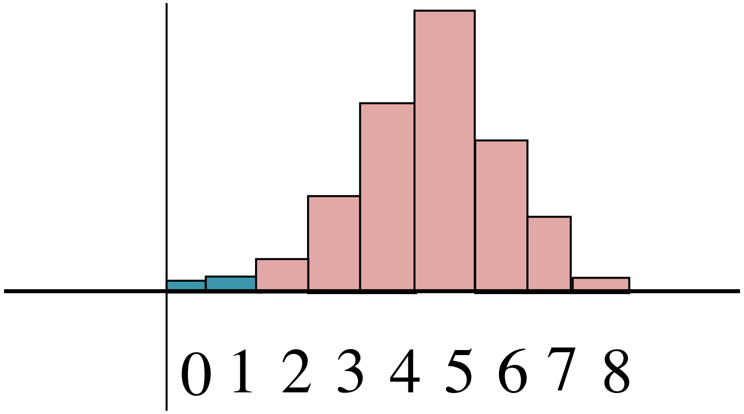
$$P(X=2)=\binom{8}{2}.0.6^2(0.4)^6$$

```
di comb(8,2)*0.6^2*0.4^6
```

```
. di binomialp(8,2,.6)
```

.04128768

Exercise



What is the probability of **2 or more cases** of disease (p=0.6) in a sample of n=8?

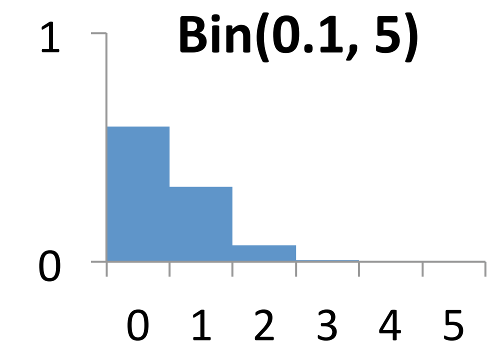
$P(X\geq 2)=P(x=2)+P(X=3)+P(X=4)+P(X=5)+P(X=6)+P(X=7)+P(X=8)$

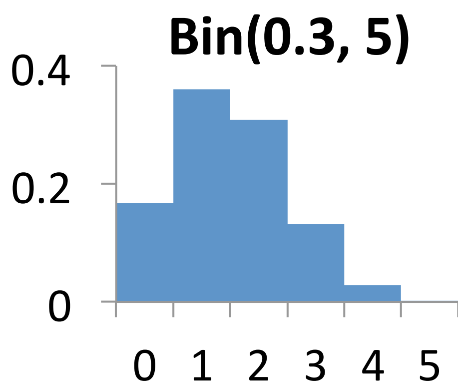
$= 1 - (P(X=0)+P(X=1))$

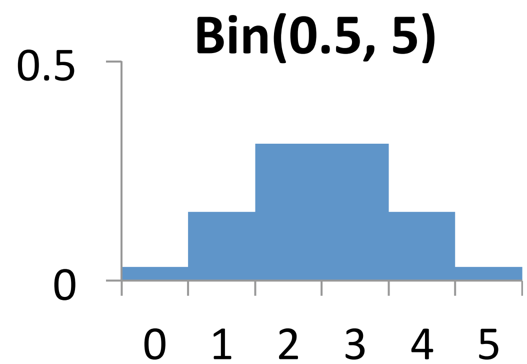
```
. di binomialtail(8,2,0.6)
.95019264
```

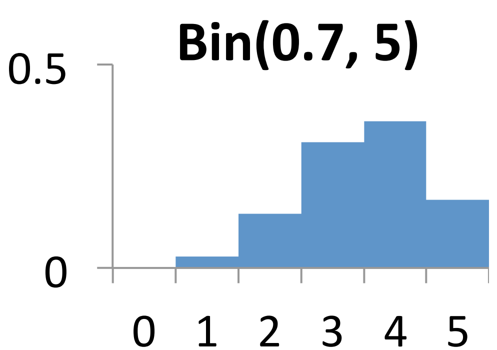
Binomial Distribution

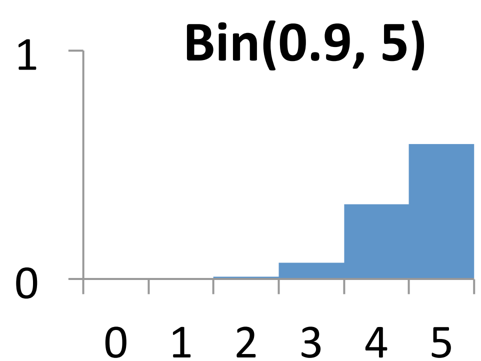
- Binomial mean = np
- Binomial variance= np(1-p)











Poisson Distribution

- Model the number of events occurring within a given time interval.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Characteristics:

1) No limit to the number of events
2) only one parameter, λ, that is the mean number of events

https://www.umass.edu/wsp/resources/poisson/

Poisson Approximation for the Binomial Distribution

When n ➔∞, p➔0, np➔λ

Binomial (n,p) ➔Poisson (λ)

$$P(X = x) = \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x} \quad \rightarrow \quad P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Normal Distribution

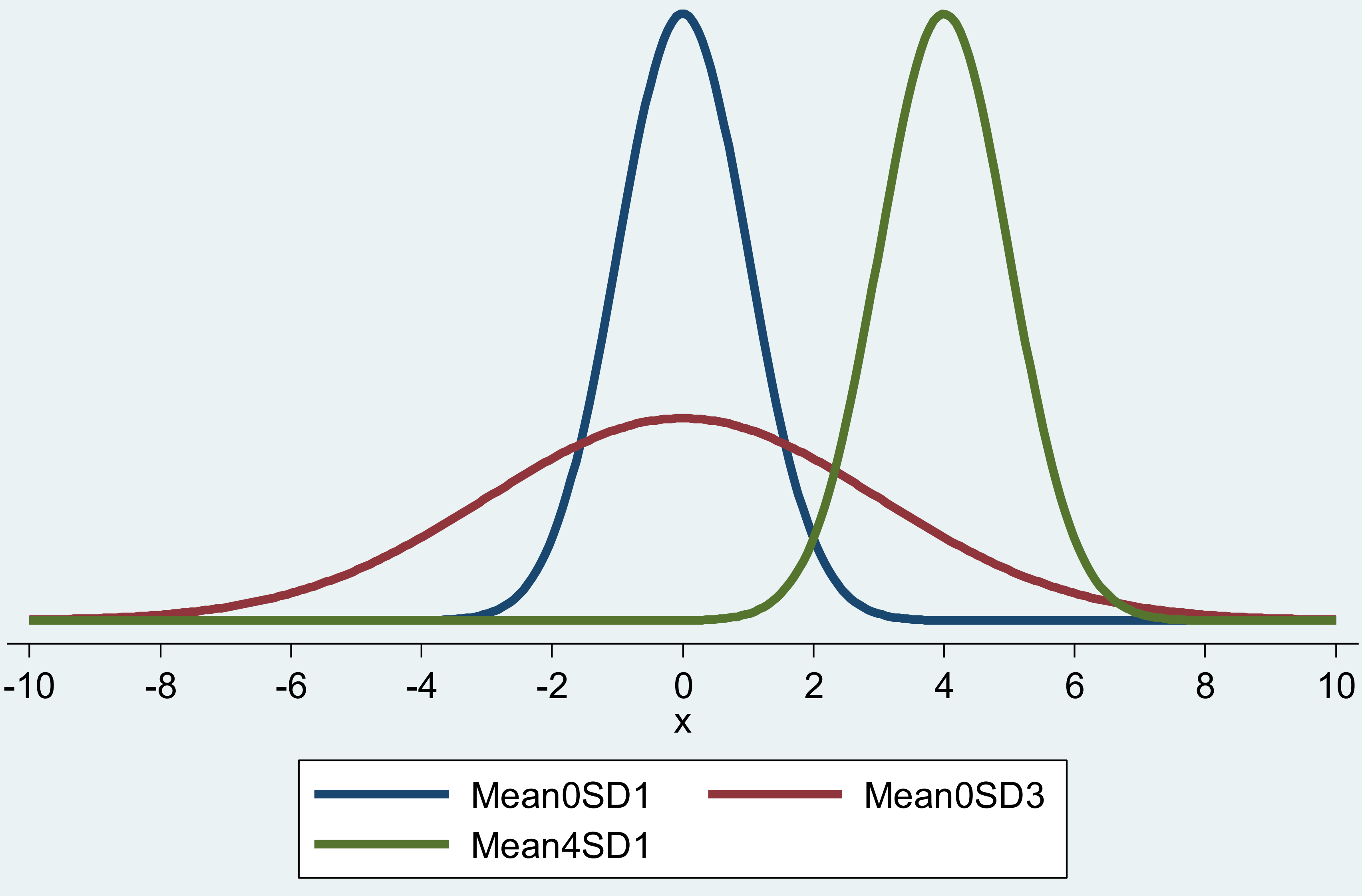
- A continuous random variable is said to be normally distributed with mean μ and variance σ^2 if its probability density function is:

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

- Normally distributed X with mean μ and SD σ:

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

Several normal distributions



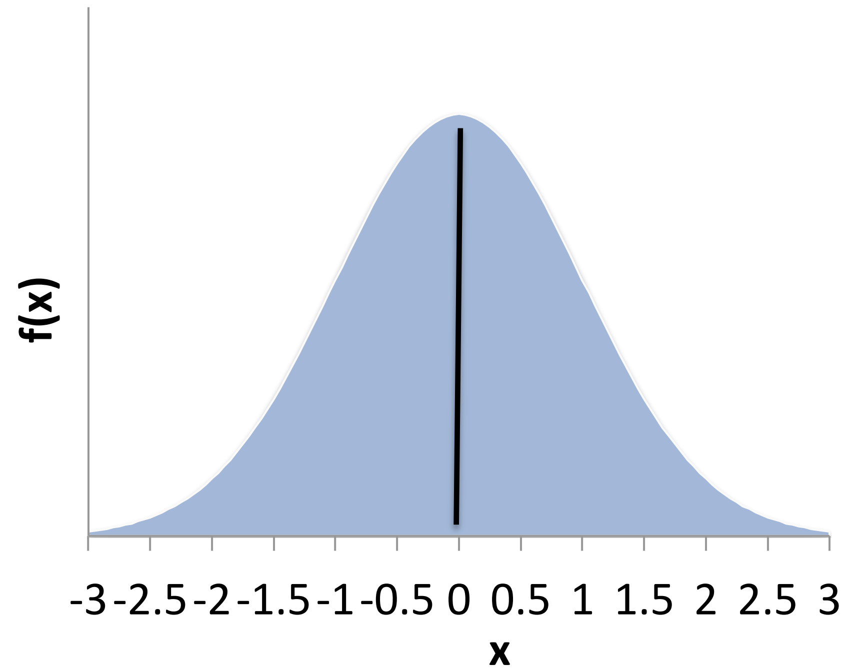
Standard Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \xrightarrow[\sigma = 1]{\mu = 0} f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Denoted: **Z ~ N(0,1)**

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

then $Z = (X - \mu)/\sigma$ is a ***standard*** normal random variable, $Z \sim N(0, 1)$.



Use Stata to calculate standard normal probabilities

The left portion of the curve P(Z<z):

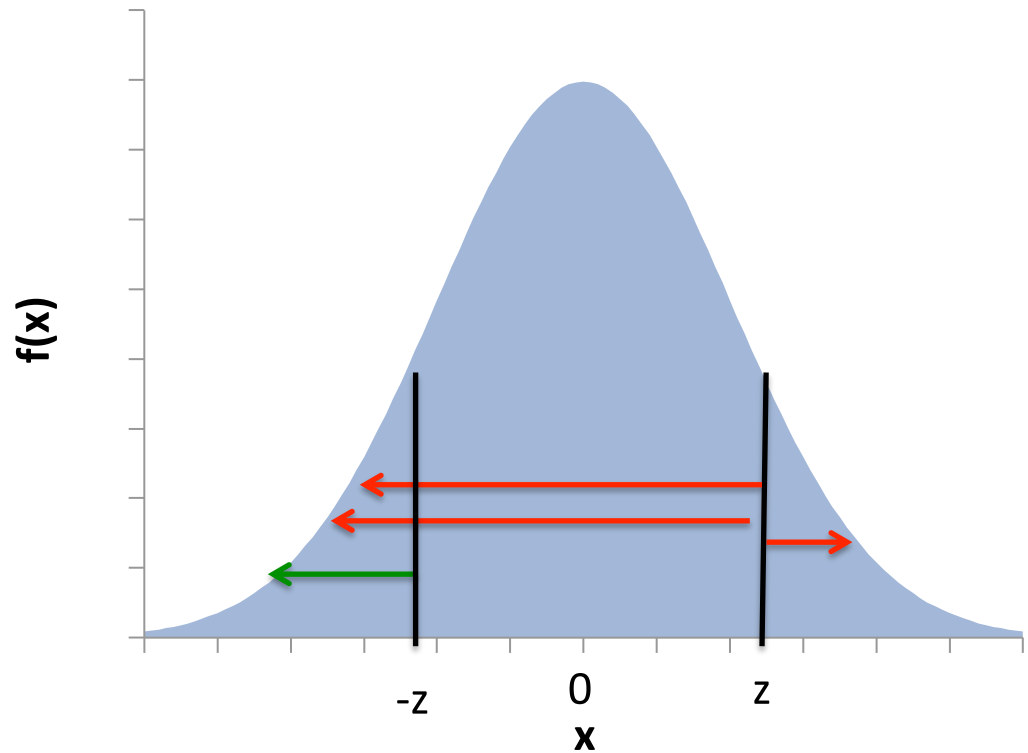
```
display normal(z)
```

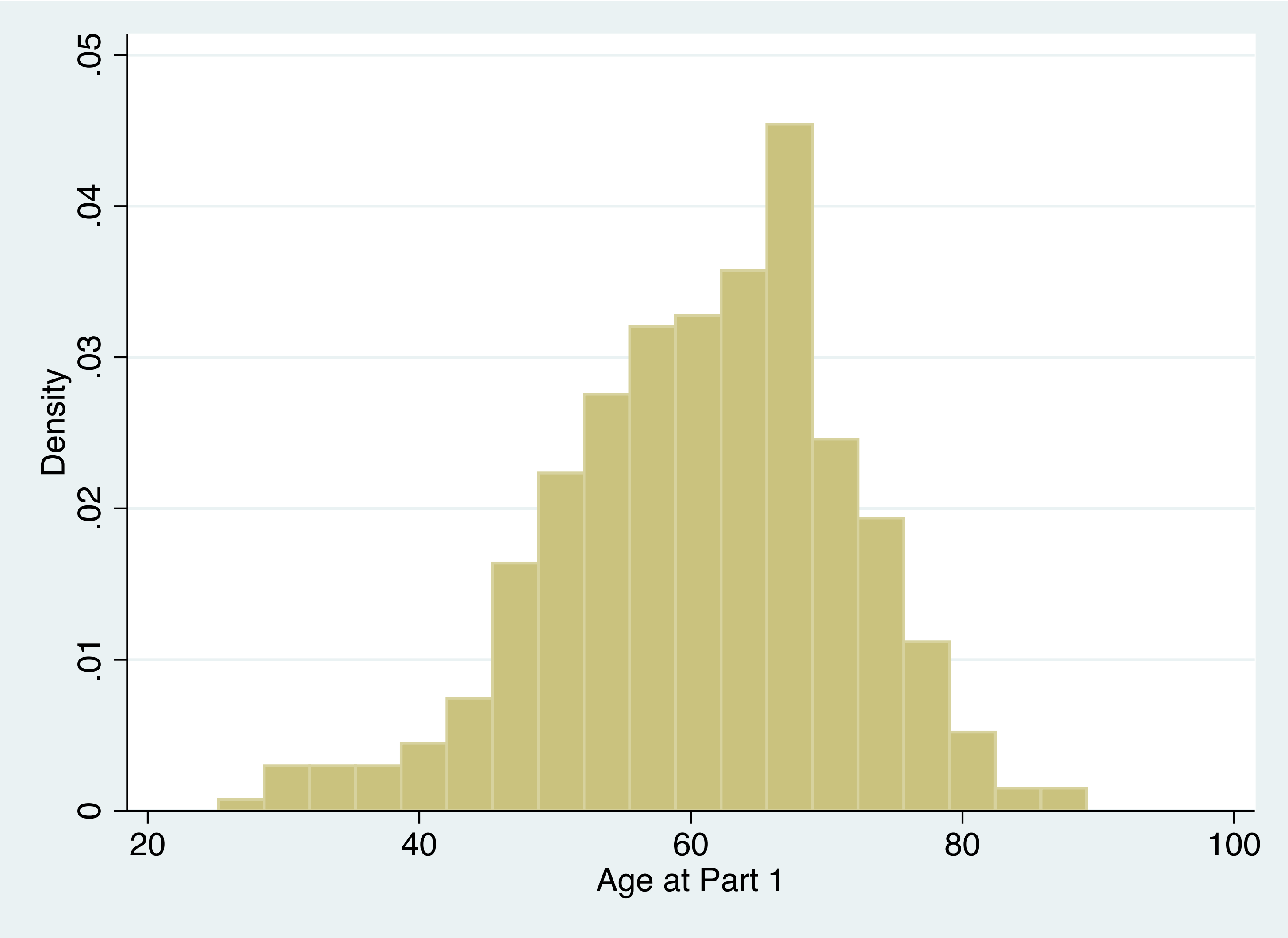
The right hand portion of the curve, P(Z≥z):

```
display 1-normal(z)
```

The middle portion:

```
display normal(z)-
normal(-z)
```





Distribution of age in CIPN cohort, N~(61, 10.7)

Q1: what's the proportion for people with age>70? (Mean=61, SD=10.7)

1) **Get the z score**
   - subtract off the mean and divide by the SD
     z=(70-61)/10.7=0.84

2) **Find P(Z>z)** using Stata

```
. di 1-normal(0.84)
.20045419
```

Q2: What is the range of age for 95% people in the CIPN cohort? (Mean=61, SD=10.7)

- Find upper and lower 2.5% values.

Lower 2.5% value:

1) Find z value satisfies $P(Z<z)=0.025$

```
. display invnormal(.025)
-1.959964
```

2) Transform back to the original age

$z=(x-61)/10.7=-1.96$

$x=-1.96*10.7+61 = 40.0$

Upper 2.5% value:

1) Find z value satisfies $P(Z \geq z)=0.025$

```
. display invnormal(.975)
1.959964
```

2) Transform back to the original age

$z=(x-61)/10.7=1.96$

$x=1.96*10.7+61 = 82.0$

So 95% people are 40 -82 years old.

If a variable that is normally distributed, the values that comprise the middle 95% or 99%:

- For the middle 95%, the interval is

  $\mu-1.96*\sigma, \ \mu+1.96*\sigma$

- For the middle 99%, the interval is

  $\mu-2.58*\sigma, \ \mu+2.58*\sigma$