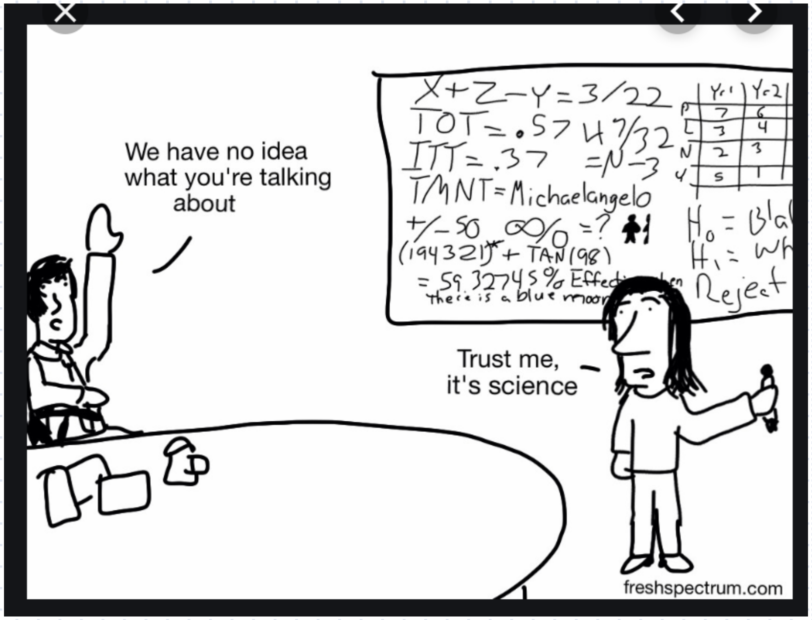
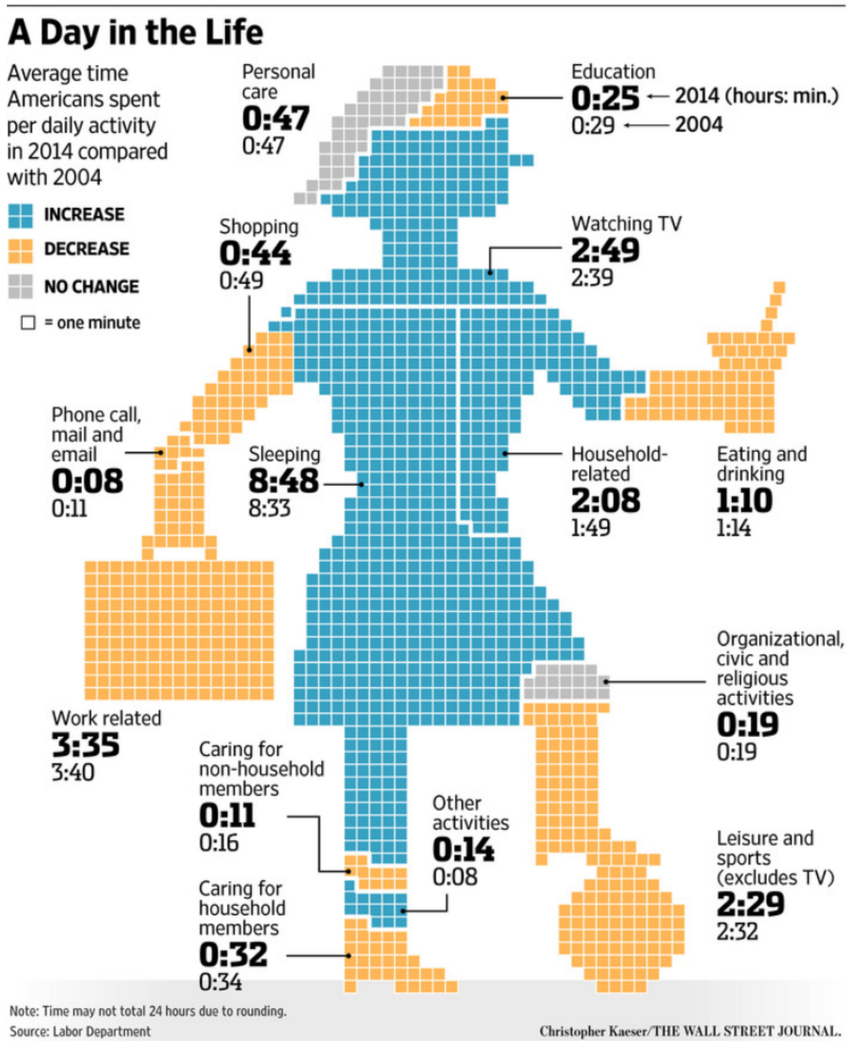
Distributions & the Central Limit Theorem 2

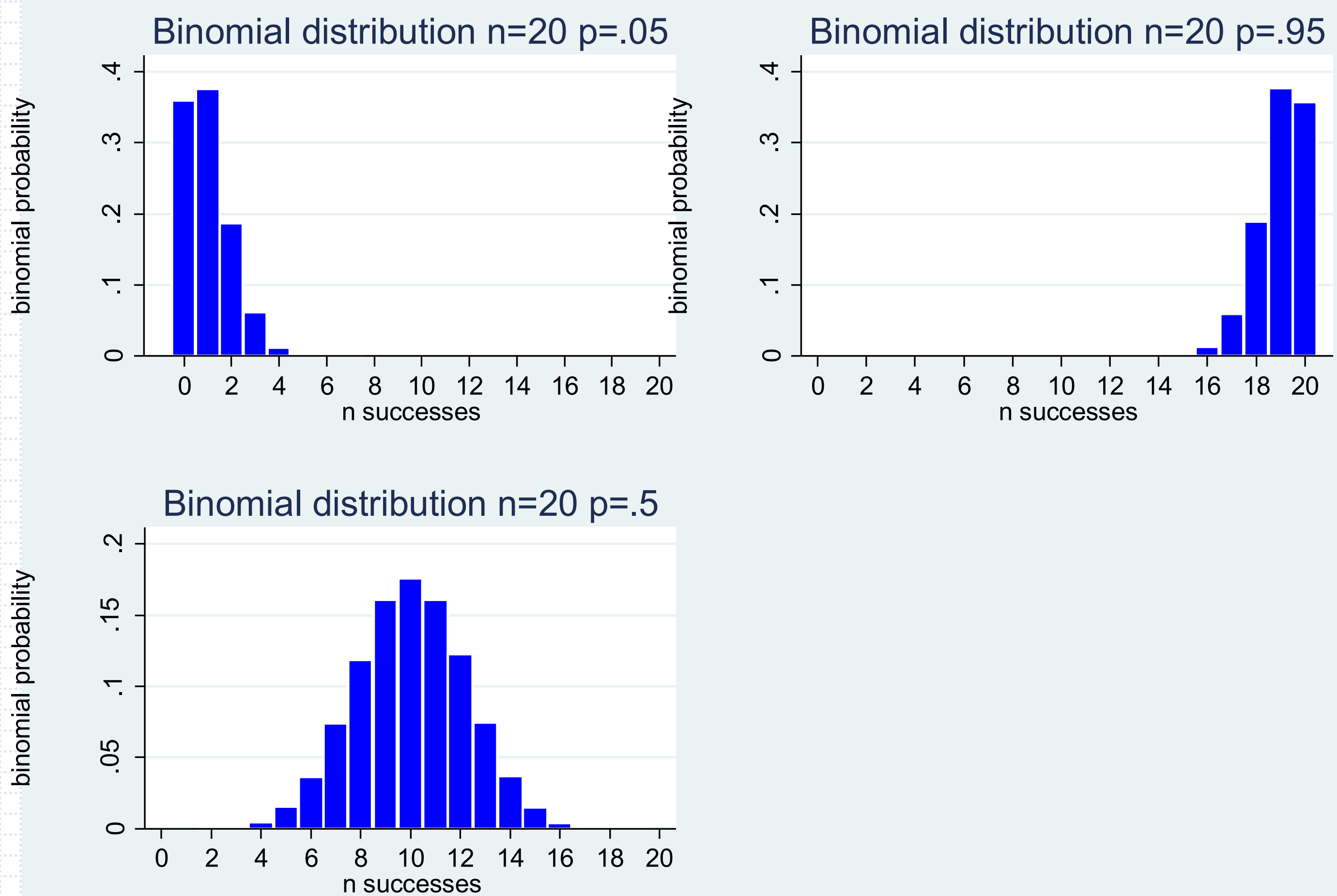


Another Bad Chart
Sufficiently Weird



Recap of binomial distribution

- The binomial distribution describes the probability of x successes in n independent trials, each with probability p of success
- The distribution will be symmetric when $p=0.5$, skewed right when $p<0.5$, and skewed left if $p>0.5$
- The possible number of "successes" will be 0 to n, because there are n "trials"
- Often "successes" are number of people with a disease, "number of trials" is the number of people in the sample


Binomial distribution n=20 p=.05
binomial probability
0 .1 .2 .3 .4
0 2 4 6 8 10 12 14 16 18 20
n successes
Binomial distribution n=20 p=.95
binomial probability
0 .1 .2 .3 .4
0 2 4 6 8 10 12 14 16 18 20
n successes
Binomial distribution n=20 p=.5
binomial probability
0 .05 .1 .15 .2
0 2 4 6 8 10 12 14 16 18 20
n successes

Recap of Binomial distribution

- $P(X=x)$ where **X** represents the random variable X that follows a binomial distribution & **x** represents what you actually get (number of successes) when you draw your sample
- In statistics, the **mean** of a theoretical probability distribution is also called the "Expected value".
- The expected value or mean of the binomial distribution is $n*p$
- For the binomial distribution, if you know the underlying probability (p) in the population, then you know that if you take your sample over and over, then the mean number of successes will be $n*p = \bar{x}$

STATA examples: Binomial

- P(X=x) = P(exactly x successes occurring) n=20,p=.05,x=2 P(X=2)

```
 ** using binomialp(n,k,p)
di binomialp(20,2,.05)
 .1886768
```

- P(X≥x) = P(of x <u>or more</u> successes) n=20,p=.05,x=2 P(X≥ 2)=1-(P(X=0)+P(X=1))

```
** using binomialp(n,k,p)
di 1 - binomialp(20,0,.05) - binomialp(20,1,.05)
 .26416048
*** using binomialtail(n,k,p)
 di binomialtail(20,2,.05)
 .26416048
```

STATA Examples: Binomial

- P(X>x) = P(More than x successes occurring)

  e.g. n=20, p=.05, x=2

  $P(X>2) = P(X \geq 3) = 1 - (P(X=0) + P(X=1) + P(X=2))$

```
** using binomialp(n,k,p)
di 1- binomialp(20,0,.05) - binomialp(20,1,.05) - binomialp(20,2,.05)
 .07548367

*** di binomialtail(n,k+1,p)
 di binomialtail(20,3,.05)
 .07548367
```

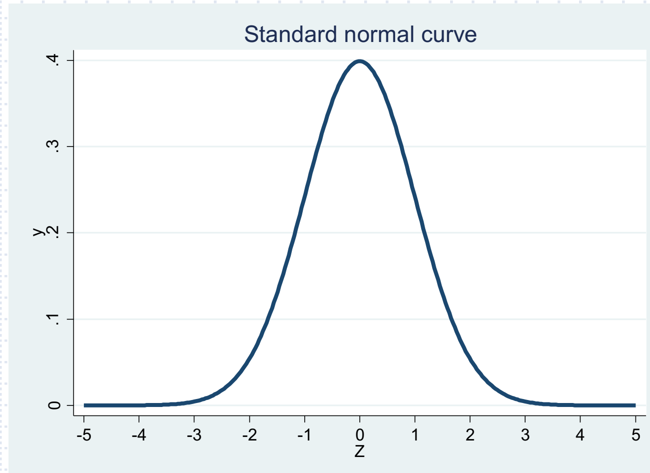
Whether you are looking at P(X≥x) vs P(X>x) matters for discrete distributions!!!

Recap from the normal distribution

- We do not calculate $P(X=x)$ or $P(Z=z)$
  - Probability of individual values are 0
- We do calculate $P(X>x)$ or $P(X<x)$ or the probability of a range of values (e.g. -1.96, 1.96)
- It does not make a difference if we use the notation $P(X>x)$ or $P(X\geq x)$ because we just said $P(X=x)=0$

STATA Normal distribution commands

Z~N(0,1) is a normal distribution with mean 0 and standard deviation 1



- P(Z > a big z) is small, e.g. P(Z>3)

```
di 1-normal(3)
.0013499
```

- P(Z < a big z) is close to 1

```
di normal(3)
.9986501
```

- But if z is negative, it is the converse

P(Z>-3)

```
di 1-normal(-3)
.9986501
```

P(Z<-3)

```
di normal(-3)
.0013499
```

Remember:

Using di normal() in Stata

for P(Z>z) use di 1-normal(z)

for P(Z<z) use di normal(z)

Recap from the normal distribution

- The normal distribution may be used to describe cutoffs for some continuous random variables with mean μ and standard deviation σ
  - We calculate **z** statistics $(\bar{x}-\mu)/\sigma$ just so we can use standard probability values
- How do you know if your data are normally distributed?
  - Histograms (stata: `hist` *`varname`*`, normal`)
  - QQ plots
  - Other statistical tests
- What to do if your data are not normal?
  - Transformations – like taking the log, or the inverse 1/x …

The theory behind means and variances

- The mean of a random variable X is the long run average of the X's and another name is the "Expected value of X".

- The variance of a random variable describes the long run spread of the possible values around the mean.

- Both are calculated based on the theoretical distribution function P(X=x) or f(x).

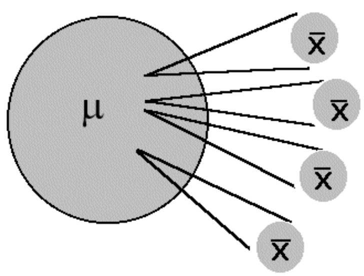
  – *Forget theory – on to taking samples!!*

Sample means and variances

- You have seen how we calculate means and variances using sample data.
- These are **estimates** of the mean and variance of the underlying population distribution from which the sample was chosen.
- When we cannot measure the entire population we take a sample
- We ***estimate*** the population characteristics, the population mean and variance, using the <u>sample</u> mean and variance
- We use ***statistical inference*** to draw conclusions about the how the estimates from the sample relate to the population values

Sampling Distributions

- Our sample must be representative of the population to make inferences
  - Random sample – each individual in the population has equal chance of being selected for the sample
  - The larger the sample, the more reliable our estimates about the population parameters will be
- Because we do not have the entire population, there is always uncertainty about our data – we could have gotten other x's in the sample
- ***Confidence intervals*** afford us a way to quantify this uncertainty

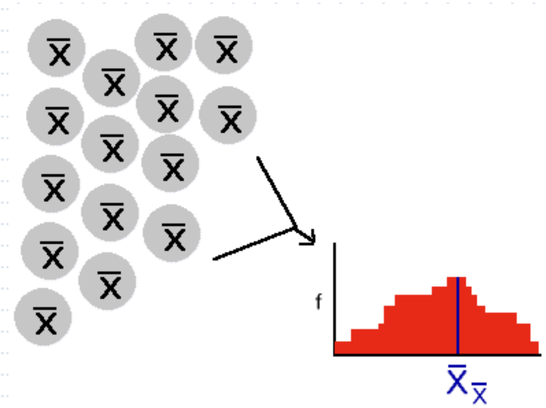


Imagine you drew many samples of size n from the population. If you did this for a long time you could then average all of them to compute your overall mean and variance. These means may vary from sample to sample but overall your mean will probably be very close to the population mean.

Sampling distributions: Example

- Treat all the $\bar{x}$ s as a sample and calculate their mean and sample variance, make a histogram to look at their distribution
- The collection of all the $\bar{x}$ s that can be obtained can be thought of as a random variable $\bar{X}$ that follows some distribution
- The distribution of all the possible $\bar{x}$ s is called the sampling distribution
- The larger the sample size within each sample is, the more the distribution of the $\bar{x}$s will look like the normal distribution.
  - Regardless of the underlying distribution of X.

*This brings us to … The Central Limit Theorem**



*Seaman J, Allen IE Staying Relevant: The Central Limit Theorem. *Quality Progress*, February, 2018.