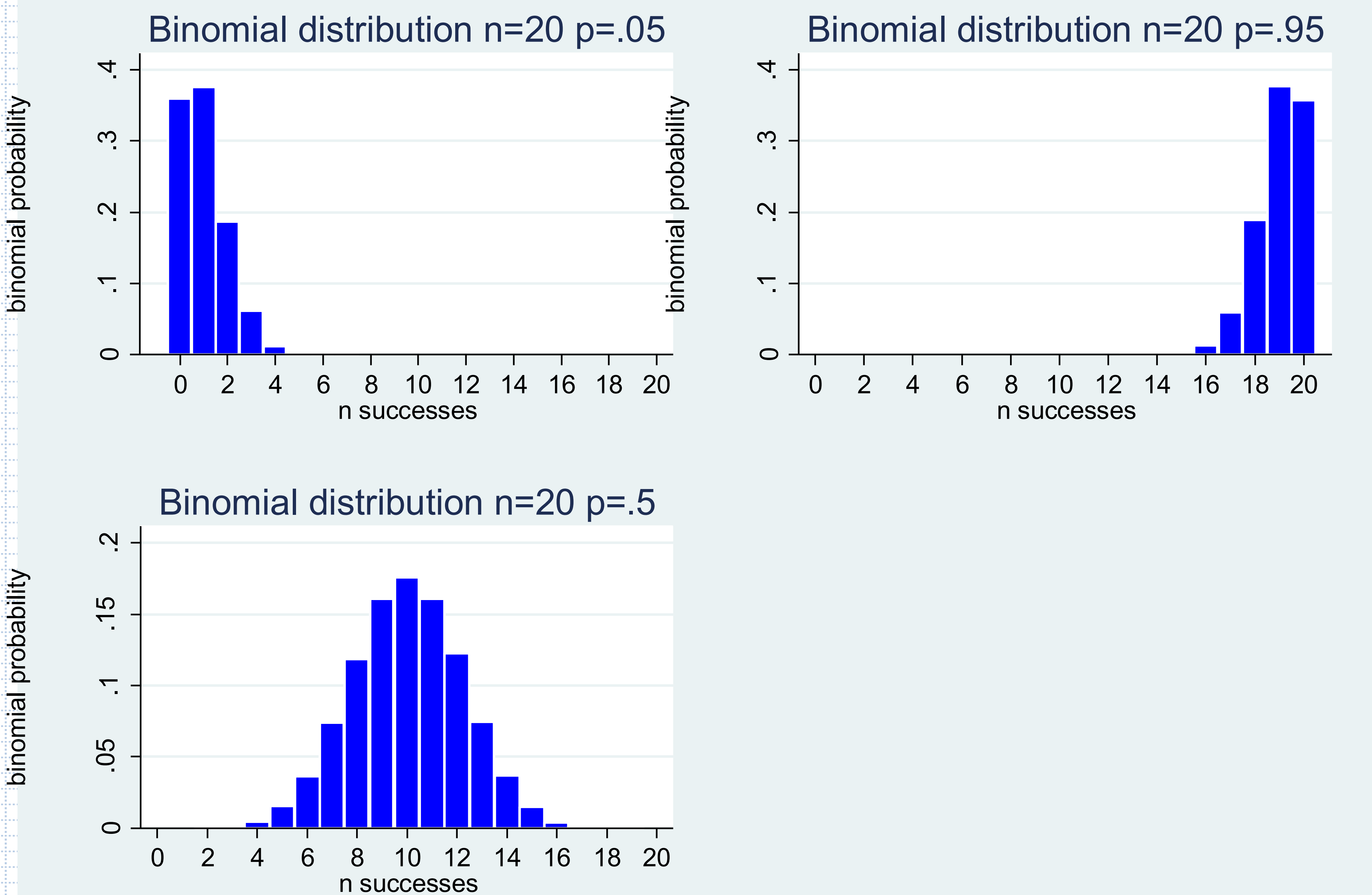
Today

- Review of binomial and normal distribution
- Sampling
- Central limit theorem
- Confidence intervals for means
- Normal approximation to the binomial distribution
- Confidence intervals for proportions

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

Binomial distribution

- P(X=x)

  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 p in the population, then you know that if you take your sample over and over, then the mean number of successes $\overline{x}$ will be n*p

Example 1: Binomial

- P(X=x) = P(exactly x successes occurring)

  e.g. n=20, p=.05, x=2 P(X=2)?

```
** using comb(n,x)*p^x*(1-p)^(n-x)
 di comb(20,2)*.05^2*.95^18
.1886768

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

Example 2: Binomial

- $P(X \geq x)$ = P(of x <u>or more</u> successes occurring)

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

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

```
** 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
```

Example 3: 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≥x) because we just said P(X=x)=0

Normal distribution

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
```

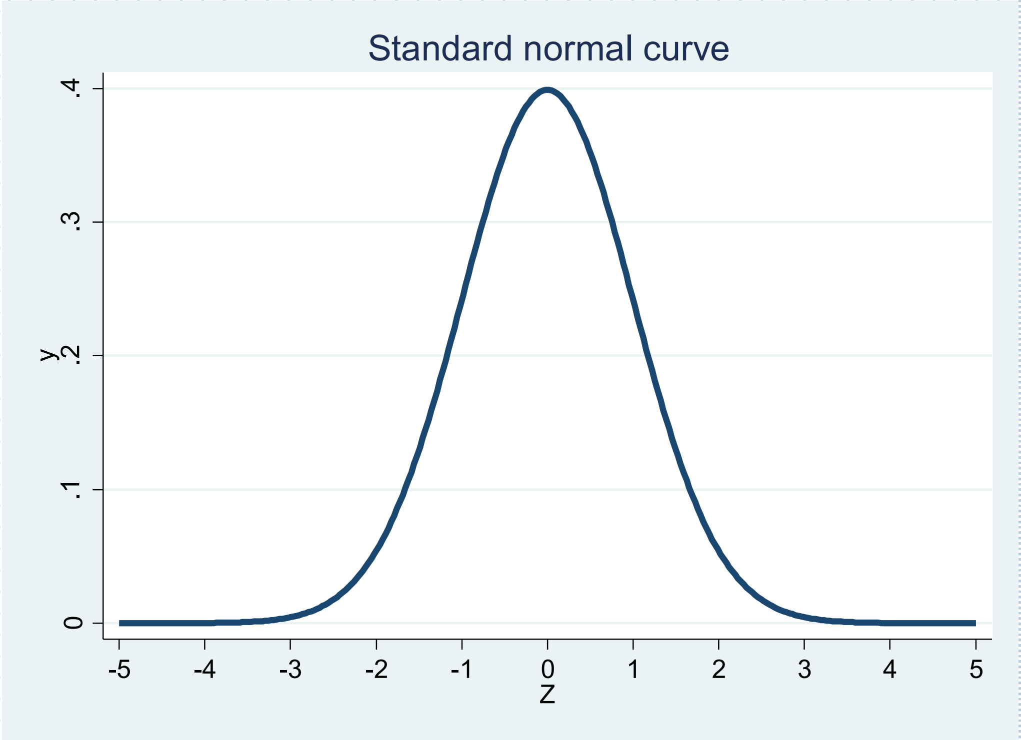
- But if z is negative, it is the converse

P(Z>-3)

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

P(Z<-3)

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



Normal distribution

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 – next Biostat class
    - Other statistical tests
- What to do if your data are not normal?
    - Transformations – like taking the log, or the inverse 1/x …

A bit more about 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 $f(x)$.

The formulae (just fyi)

- The general formula for the mean of discrete distributions is:

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

For a binomial x_i is the number of successes (0,1,2, …n), and $P(X= x_i)$ is the height of each bar (obtained using the binomial formula), and it works out to equal np

- For continuous distributions the formula for the mean is

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

For the normal distribution this works out to be the formula μ

- The formula for the variance of a random variable X is $E[(X-\mu_x)^2]$

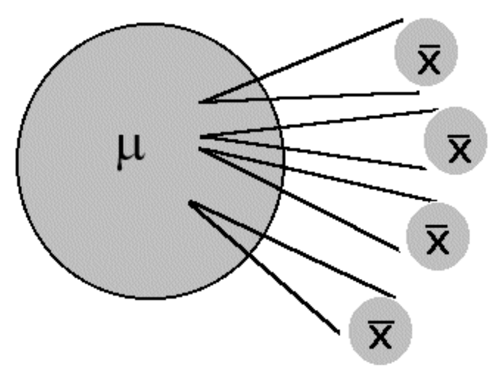
Sample means and variances

- The formulae presented in the first lecture are used for calculating sample means and variances using sample data.

- They are **estimates** of the mean and variance of the underlying population distribution from which the sample was chosen.

Sampling

- When we cannot measure the entire population we take a sample
- We *estimate* the population characteristics, i.e. the mean and variance of our data, using the sample mean and variance
- We use **statistical inference** to draw conclusions about the how the estimates from the sample relate to the population values



Timothy Shortell
http://www.shortell.org/book/chap10.html

Sampling

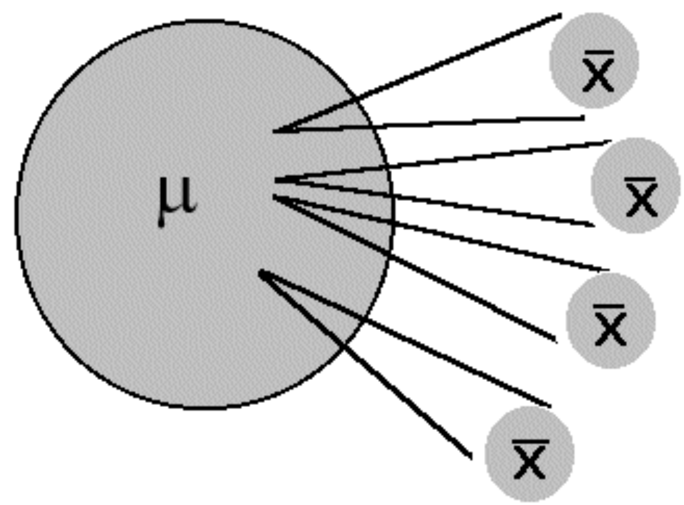
- Our sample must be representative of the population to make inferences
  - Random sample – each individual in the population has equal chance of being selected for the sample
  - The larger the sample, the more reliable our estimates about the population parameters will be
- Because we do not have 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

Sampling distributions: Example

- Imagine you drew a sample of size n from a population and measured a random variable X, say systolic blood pressure and calculated the sample mean, $\bar{x}_1$ from the x_i
- Then you drew another sample of size n, and calculated $\bar{x}_2$
- If you repeat for a long time (say mmm times) you will have a large collection $\bar{x}_i$ s generated from the samples of size n $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{\text{mmm}})$

Sampling distributions: Example

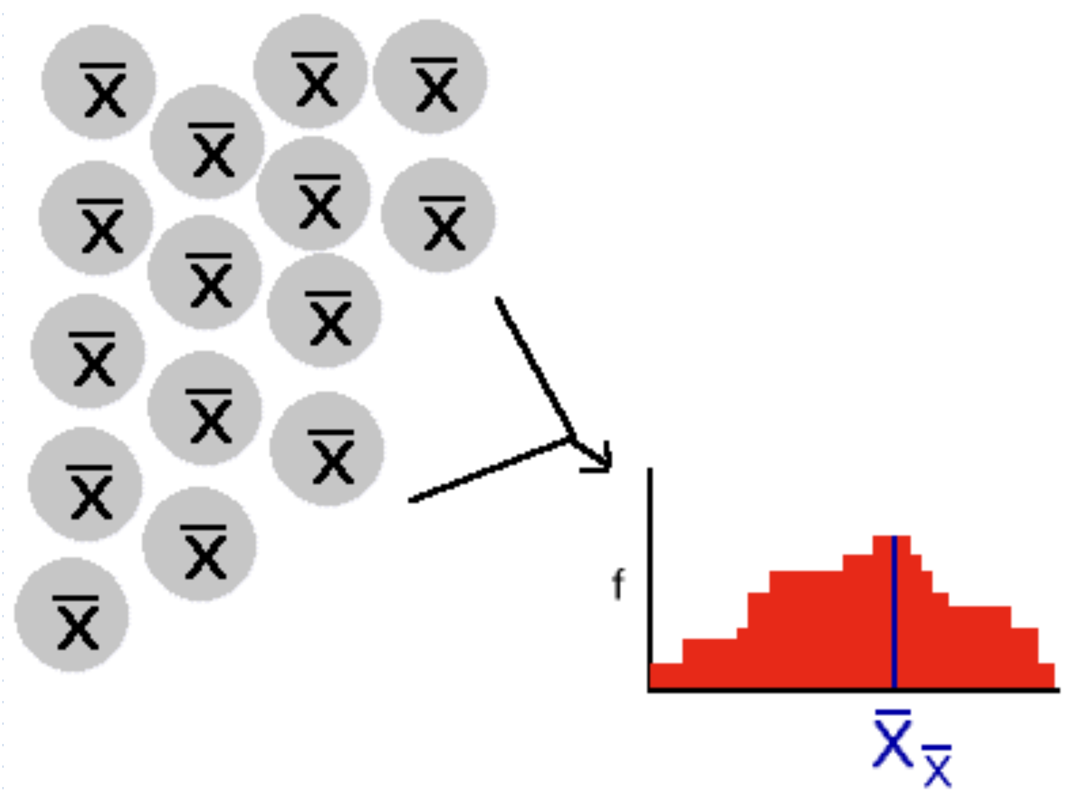
- The $\bar{x}$ s will differ from sample to sample due to sampling variability – each sample will most likely be different



Timothy Shortell
http://www.shortell.org/book/chap10.html

Sampling distributions: Example

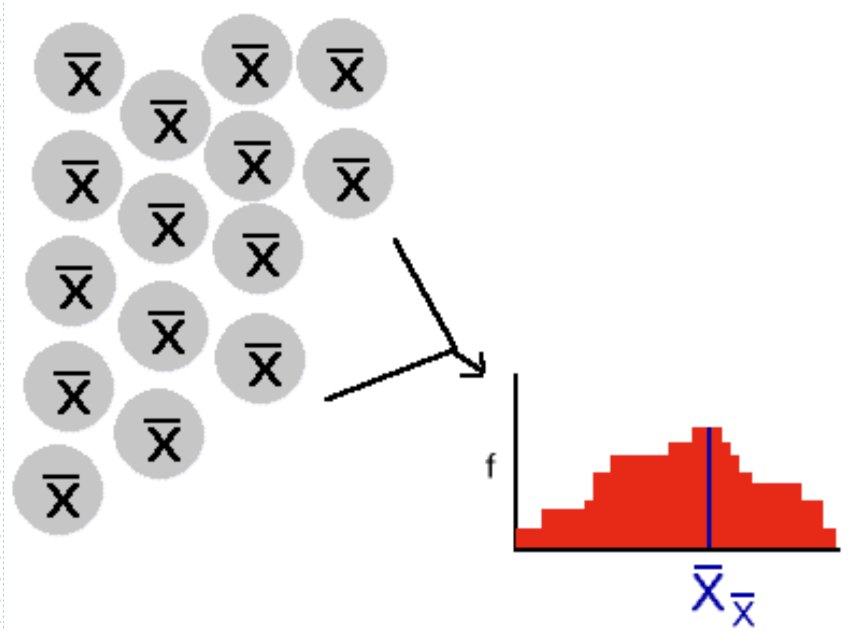
- Treat all the $\bar{x}$ s as a sample and calculate their mean and sample variance, make a histogram to look at their distribution



Timothy Shortell
http://www.shortell.org/book/chap10.html

Sampling distributions: Example

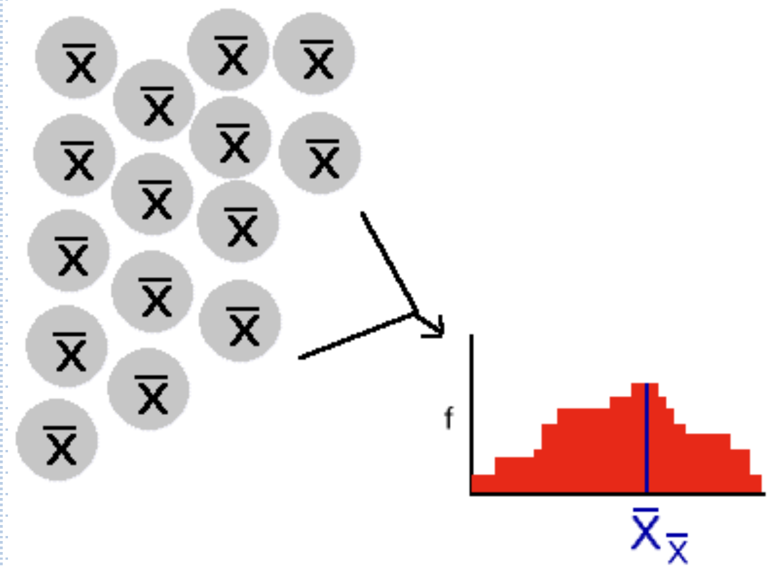
- The collection of all the $\bar{x}$ s that can be obtained can be thought of as a random variable $\bar{X}$ that follows some distribution
- The distribution of all the possible $\bar{x}$ s is called the <u>sampling distribution</u>



Timothy Shortell
http://www.shortell.org/book/chap10.html

Sampling distributions: Example

- 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.



- Do you take multiple samples?

Timothy Shortell
http://www.shortell.org/book/chap10.html

Central limit theorem

- If you have a random variable that comes from a distribution with mean=μ and standard deviation=σ, the following is true for the sampling distribution (i.e. the distribution of all possible sample means) from samples of size *n*
  – If *n* is large enough, the shape of the sampling distribution will be approximately normal

  – The mean of the sampling distribution is μ

  – The standard deviation of the sampling distribution is $\sigma/\sqrt{n}$

Central limit theorem

- If we take a sample from any distribution (could be skewed, or discrete, or whatever) of size n, and take the mean, and repeat this over and over, the distribution of the means will be normally distributed with mean=the original distribution mean μ and standard deviation= $\sigma/\sqrt{n}$, *if n is large enough*

- The more symmetric the distribution of the underlying data (not the sample means), the smaller the n needed for the distribution of $\bar{x}$ to become normal-like

CLT: Why do we care?

- If we know that the distribution of $\bar{X}$ is normal, then we can use that to calculate the probability of obtaining an $\bar{X}$ as extreme as the one we did (i.e. that we got $\bar{x}$ as our sample mean)

- I.e., we can find things like P($\bar{X} \geq \bar{x}$) because $\bar{X}$ follows a normal distribution

- We can compute the standard normal (z-scores) from this.

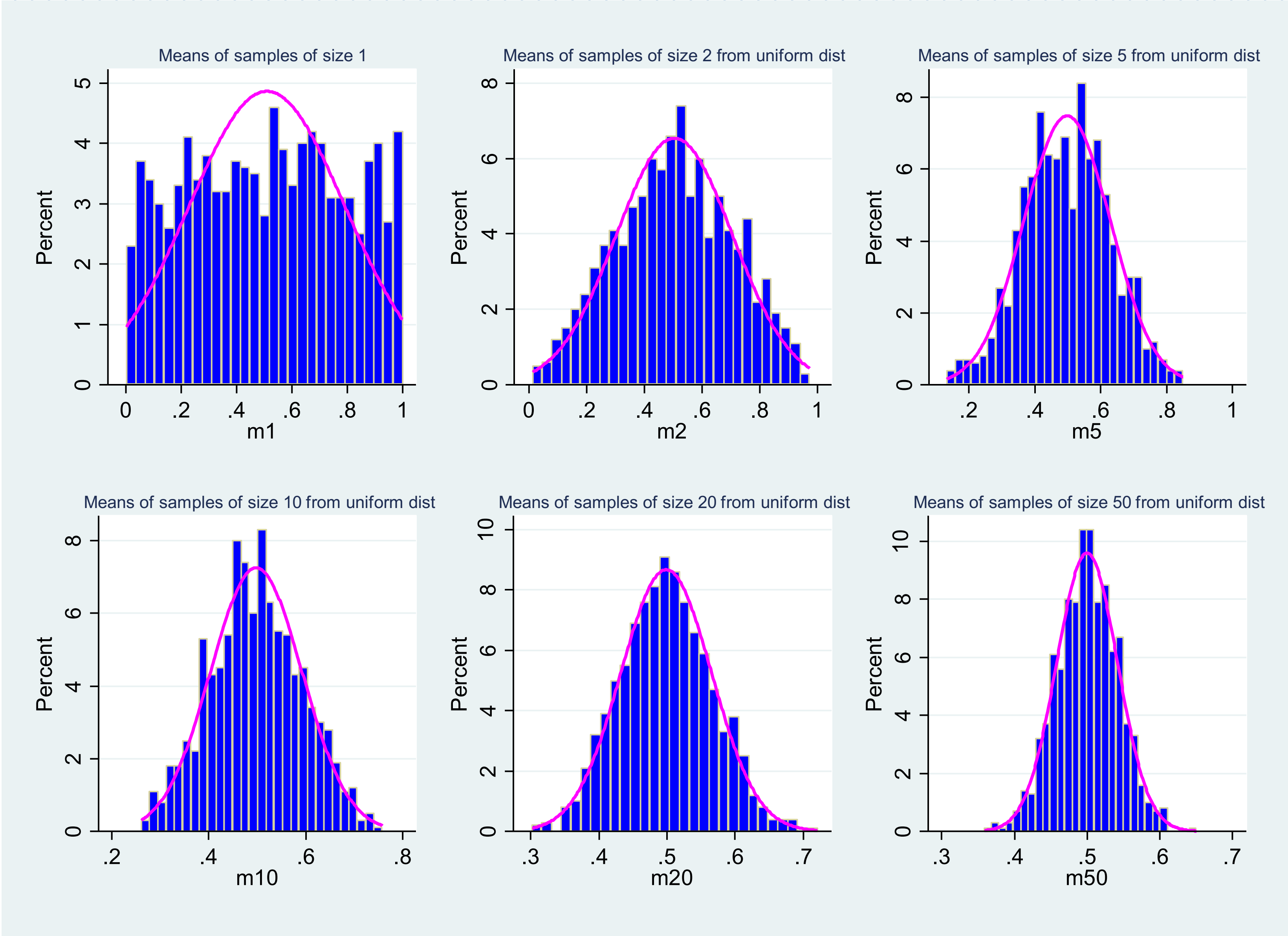
Sampling distributions

- Does this make sense?
  - It makes sense that the distribution of means would cluster around the population mean
  - It makes sense that the variability in the means is smaller than in the raw data because the extreme values are already averaged out (remember $\sigma/\sqrt{n}$)
  - The part about the distribution being normal if n is large enough? Mathematical proof … But we can demonstrate it for several examples

Sampling distributions
Distribution of the Sample Mean

- σ is the standard deviation of the original distribution
- $\sigma/\sqrt{n}$ is called the *standard error,* or more precisely, the *standard error of the mean*, and it is the standard deviation of the distribution of the sample mean.

Distributions of the means of Uniformly distributed random variables



Uniform Distribution Example

- Taking samples from the following Uniform distribution
- Uniform(1, 100) with different sample sizes
- http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html

Using the Central Limit Theorem

- Suppose we sampled from a HIV-infected population with mean μ CD4 count = 250 cells/mm^3 and standard deviation σ = 200 cells/mm^3.
- If we select repeated (a lot) samples of size 50, what proportion of the samples will have a mean value of less than 100 cells/mm^3 ?
- Using the CLT, we know that the mean of each of the samples, $\overline{x}$ is itself a random variable, that follows a normal distribution with mean μ=250 and standard error $\sigma/\sqrt{n}=200/\sqrt{50}$

```
. di 200/sqrt(50)
28.284271
```

Using the Central Limit Theorem

- So $\bar{X} \sim N(250, 28.3)$
- Then we know that $(\bar{x}-250)/(200/\sqrt{50}) \sim N(0,1)$
- We wanted to know what proportion of the means would have a value of <100, we want $P(\bar{X}<100)$ then

  $z=(100-250)/(200/\sqrt{50})$

  $= -150/28.3 = -5.3$

$P(Z<-5.3) = ?$

Using the Central Limit Theorem

- What level of CD4 count is the <u>lower</u> 2.5th percentile of the mean values?
- For what value of z is P(Z<z)=.025 ?

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

- Now we use this to get back to $\bar{X}$, using

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- So -1.96=($\bar{x}$-250)/(200/$\sqrt{50}$)
- Rearranging, $\bar{x}$= -1.96*200/$\sqrt{50}$+250

```
di -1.96*200/sqrt(50)+250
194.56283
```

Using the Central Limit Theorem

- What level of CD4 count is the upper 2.5th percentile of the mean values?
- What is the z-value that solves P(Z>z)=0.025?
- Remember invnormal gives use the z-value for P(Z<z)=p, so we need to use 1-p in the invnormal command

```
di invnormal(1-.025)
1.959964
```

- So 1.96=($\bar{x}$-250)/(200)/$\sqrt{50}$)
- Rearrange to solve for $\bar{x}$

```
di 1.96*200/sqrt(50)+250
305.43717
```

Using the Central Limit Theorem

- Now we have the lower and upper 2.5% percentiles <u>of the distribution of the sample means</u>.
- The interior area contains 95% of the sample means.
- 95% of the means from samples of size 50 that come from the underlying distribution ~N(250,200) will lie within this interval (194.6, 305.4)

Using the Central Limit Theorem

- If we selected just one sample of size 50 (what you usually do) and the sample mean was outside these limits (e.g. 315), it possibly came from a different underlying population, or that a rare (<5% probability) event had occurred.
  - Because we had said that 95% of the time the sample mean will be in the range of 195-305
  - We could say this because the central limit theorem told us that the distribution of the sample means is approximately a normal distribution with mean 250 and standard deviation 200/ $\sqrt{50}$

- This interval for the mean depends on the sample size, n. If the sample size was 300, what would be the interval?
- Lower limit: $-1.96 = (\bar{x} - 250)/(200/\sqrt{300})$

```
. di -1.96*200/sqrt(300) + 250
227.36787
```

- Upper limit $1.96 = (\bar{x} - 250)/(200/\sqrt{300})$

```
. di 1.96*200/sqrt(300) + 250
272.63213
```

- The lower and upper limits would be:

  $227.4 <= \bar{X} <= 272.6$

Which are narrower than the limits for n=50

  (194.6, 305.4)

- As n increases, the width of the interval decreases

Confidence intervals for means

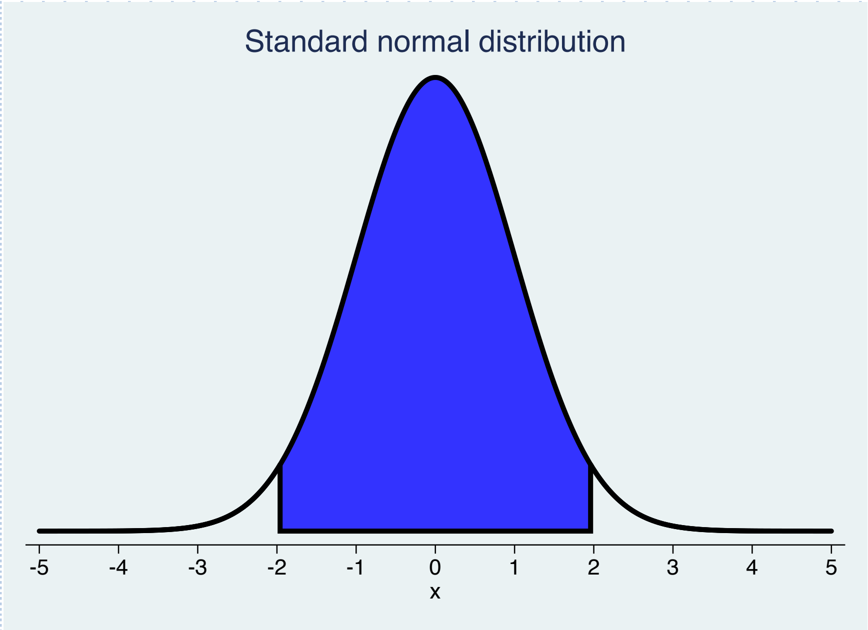
- $\bar{X}$, the sample mean, is a *point* estimate of μ, the population mean
- Different samples will yield different $\bar{X}$*s,* so we cannot be certain how our estimate differs from μ
- *Interval estimation* provides a range of reasonable values that contain the population parameter (in this case μ) with a certain degree of confidence
- This interval is called a *confidence interval*

Confidence intervals for means

- We put together what we learned about the normal distribution and the central limit theorem in order to construct confidence intervals
- By the CLT, $\bar{x}$ follows a normal distribution if n is sufficiently large $\bar{X}$ ~ $N(\mu,\sigma/\sqrt{n})$
- So, $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ follows a standard normal distribution $Z \sim N(0,1)$

Confidence intervals for means

- We use what we know from the CLT to calculate confidence intervals for means in general
- We know from examining the standard normal distribution that $P(-1.96 \leq Z \leq 1.96) = 0.95$



Confidence intervals for means

- P(-1.96 ≤ Z ≤ 1.96) = 0.95

- We also know by the CLT that $\bar{X}$ ~ *N(μ, σ /√n)* so $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

- Substituting the formula for Z into the above we get

$$P(-1.96 \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le 1.96) = 0.95$$

- Rearranging and multiplying by -1 within the parentheses we get:

$$P(\bar{X} - 1.96\sigma / \sqrt{n} \le \mu \le \bar{X} + 1.96\sigma / \sqrt{n}) = 0.95$$

Confidence intervals for means

Thus the lower 95% confidence limit for μ is $\bar{X} - 1.96\sigma / \sqrt{n}$

And the upper 95% confidence limit for μ is $\bar{X} + 1.96\sigma / \sqrt{n}$

We say we are 95% confident that the interval we calculate using the above formulae includes ***μ*** (the true mean of the distribution)

Confidence intervals for means

- $\bar{X}$ is a random variable that is the result of sampling

- μ is a population parameter that is fixed in perpetuity; it as the same value irrespective of the sample

- μ is either in the interval you calculate or it is not

- **<u>What is random is the interval</u> because it is based on the sample**

Confidence intervals for means

- So we say that the probability that the interval contains the true population mean is 95%

- If we were to select 100 random samples from the the population and calculate confidence intervals for each, approximately 95 of them would include the true population mean μ (and 5 would not)

Confidence intervals for means

- 90% confidence interval
  - Replace 1.96 in the formula with 1.64
- 99% confidence interval
  - Replace 1.96 in the interval with 2.58

Generic formula: $$(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$$

Where 100%*(1-α) is the % of the confidence interval

E.g. for a 95% confidence interval, α =0.05, and we use $z_{0.025}$ =1.96

Confidence intervals for means

- How to get a tighter interval?
  - Decrease the confidence level

Confidence level	α	$Z_{\alpha/2}$
99%	.01	2.58
95%	.05	1.96
90%	.10	1.64
80%	.20	1.28

Confidence intervals for means

- How to get a tighter interval?
  - Increase n

n	95% confidence limits	Length of interval
10	$\bar{x} \pm 1.96\sigma/\surd 10 = \bar{x} \pm 0.620\sigma$	1.240σ
100	$\bar{x} \pm 1.96\sigma/\surd 100 = \bar{x} \pm 0.3920\sigma$	0.784σ
1000	$\bar{x} \pm 1.96\sigma/\surd 1000 = \bar{x} \pm 0.062\sigma$	0.124σ

Confidence intervals for means

- What to do when σ is not known? (In practice, always)
- By the Central Limit Theorem, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ follows a normal distribution, if n is sufficiently large
- Can we substitute s, the sample standard deviation for σ ?
- s is not a reliable estimate of σ if n is small

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

Confidence intervals for means

- If X is normally distributed, and a sample of size n is chosen, then $t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$ follows a Student's t distribution

  with n-1 degrees of freedom

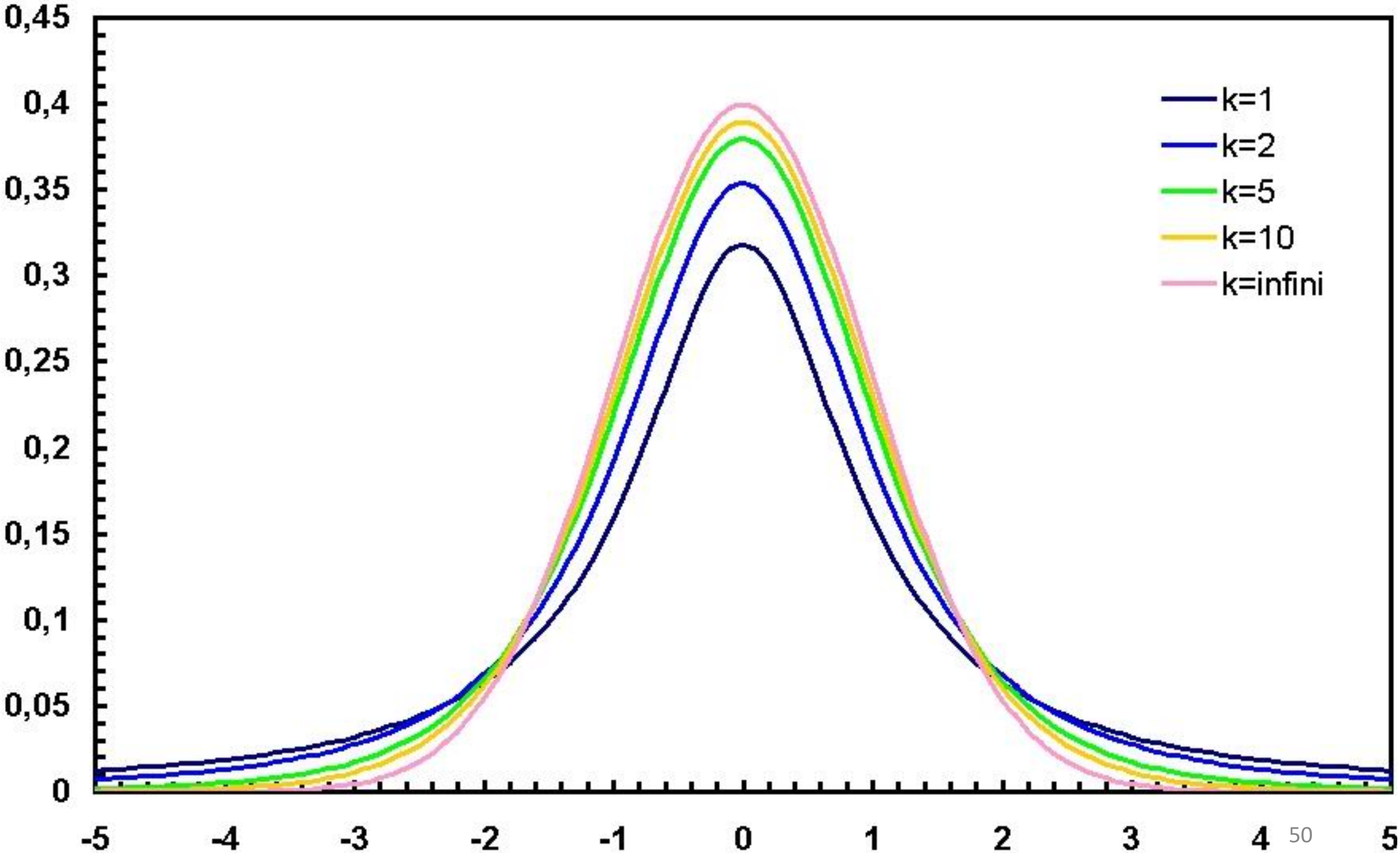
Who was Student?

- This is denoted t_{n-1}

Student's t distribution

- The mean of the t distribution is 0 and the standard deviation is 1
- The t distribution is symmetric and bell-shaped, but has heavier tails than the standard normal – extreme values are more likely to occur
- For small n, the tails are fatter
- For large n, the t distribution approaches (i.e. becomes indistinguishable from) the standard normal distribution

The t-distribution



Student's t-distribution

- There are separate curves for each degree of freedom (df)

  Most basic stat books still include tables of the t-distribution

- Better to use Stata:
- P(T≥t) is calculated using ttail *****

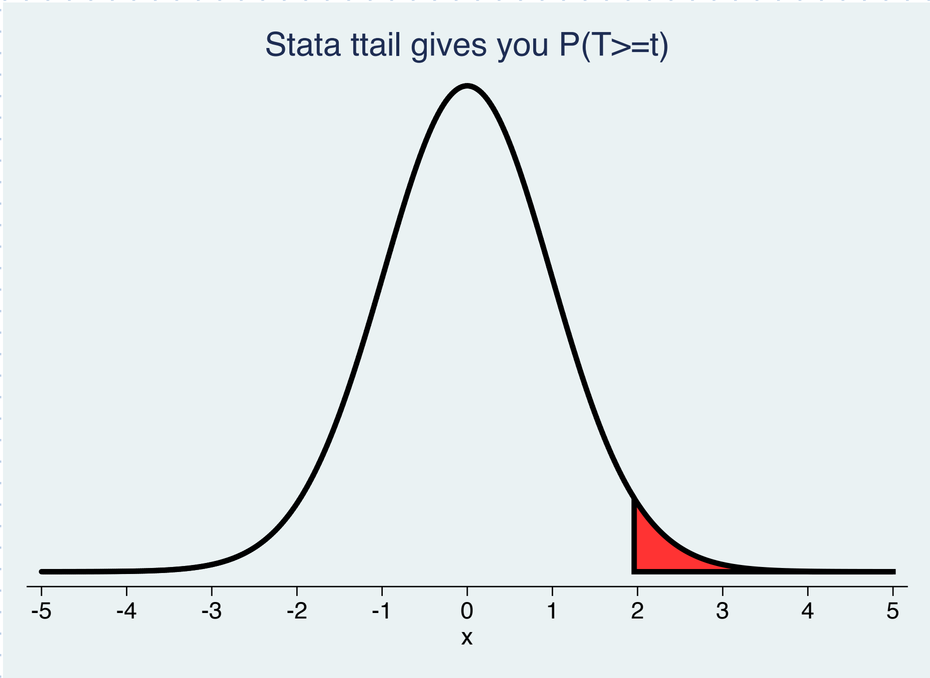
  ***** ***note that normal() gives <u>P(Z<z)</u>!!!***

- The code is ttail(df,t)
- E.g., P(T>1.96) n=20

```
display ttail(19,1.96)
.03241449
```

<u>USE n-1 for the df</u>

<u>What are degrees of freedom?</u>



Student's t-distribution

- To find the value for which P(T>t)=p use invttail(df,p) in Stata
- For example, for what t is P(T>t)=.025 for a sample of size 20?
- The answer for this t cutoff value is denoted $t_{19,.05}$

```
display invttail(19,.025)
2.0930241
```

Student's t-distribution

95% confidence intervals for means when σ is not known

- So using the t-distribution, the formula for a for a 95% confidence interval for a mean is:

$$(\bar{X} - t_{df,0.025} s/\sqrt{n}, \bar{X} + t_{df,0.025} s/\sqrt{n})$$

Where df =n-1

Confidence intervals for means

- Remember that when n is large, the t distribution approaches the normal distribution
  - E.g. $z_{0.025} = 1.960$
  - While $t_{n-1,0.025} = \rightarrow$

n	$t_{n-1,0.025}$
2	12.706
3	4.303
5	2.776
10	2.262
50	2.010
100	1.984
200	1.972
300	1.968
500	1.965
1000	1.962
1500	1.962

Confidence intervals for means: Example

- CD4 cell count among HIV positives diagnosed at Mulago Hospital
  - N=999
  - Sample mean = 329.2
  - Sample SD = 266.1
  - t cutoff?

```
. di invttail(998,.025)
1.9623438
```

  - 95% CI

    = (329.2 – 1.962*266.1/√999, 329.2 + 1.962*266.1/√999)

    = (312.7-345.7)

Confidence intervals for means in Stata

```
. . summ cd4count

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    cd4count |       999    329.2332    266.1177          1       1932


. mean cd4count

Mean estimation                     Number of obs    =     999

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    cd4count |   329.2332   8.419592      312.7111    345.7554
--------------------------------------------------------------


. ci mean cd4count

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    cd4count |        999    329.2332    8.419592        312.7111    345.7554
```

Confidence intervals for means in Stata

- Note that some statistical output gives you the SE or the SEM, which stands for standard error or standard error of the mean.
- This is s/$\sqrt{n}$ which is what you will use in confidence intervals

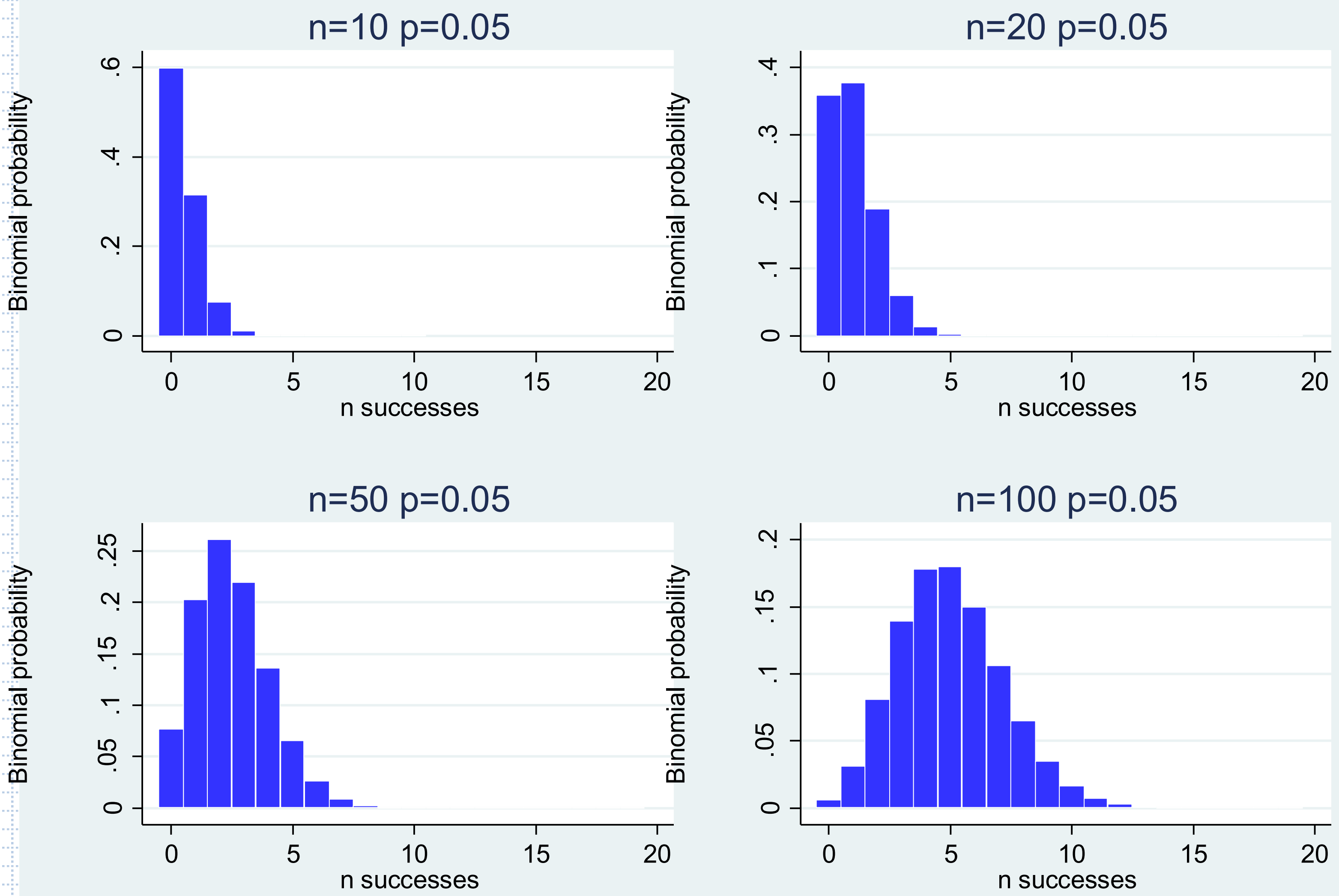
Normal approximation to the Binomial distribution

- Remember that binomial distributions are used to describe the number of "successes" in n trials P(X=x)
- The parameters of the binomial distribution are n and p, and the mean=np and standard deviation σ=$\sqrt{np(1-p)}$

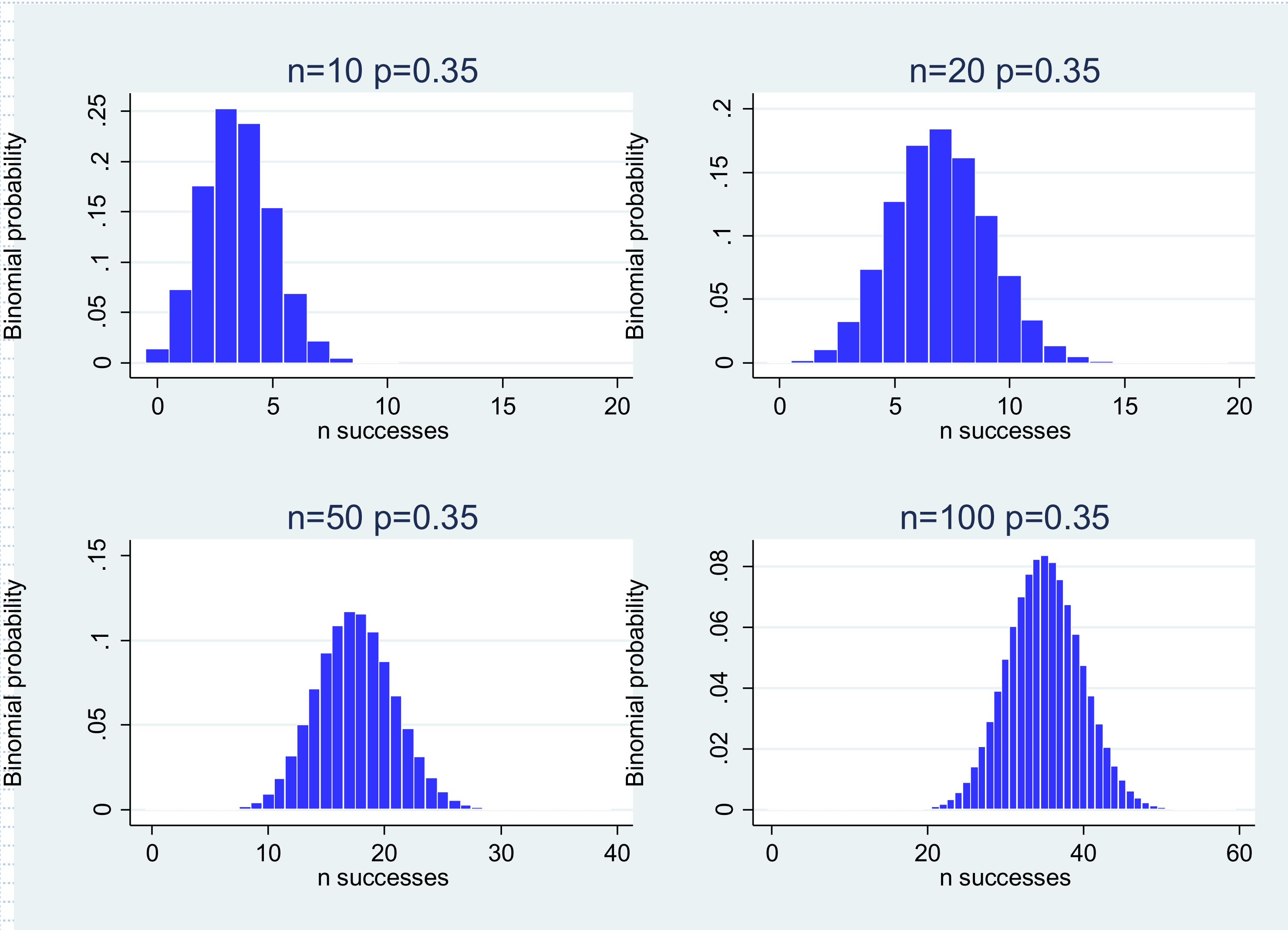
Normal approximation to the Binomial distribution

- As n, the number of “trials”, increases, the binomial distribution more closely resembles the normal distribution
- Note that the binomial distribution approaches normality at smaller sample sizes when p is closer to 0.5

Normal approximation to the Binomial distribution



Normal approximation to the Binomial distribution



Normal approximation to the Binomial distribution

- Therefore you could use the normal distribution to look up the probability of observing X or more (or less) successes
  - You would use n*p as the mean
  - You would use np(1-p) as the variance
  - You would use sqrt[np(1-p)] as the standard deviation

Normal approximation to the Binomial distribution

- What is the probability of 30 or more successes in a sample of size 50 where p=0.45?
- Using the binomial distribution for P(X>=30; n=50, p=.45)

```
. di binomialtail(50,30,.45)
.02353582
```

Normal approximation to the Binomial distribution

- Using the normal approximation
  - Mean=n*p= 50*.45=22.5
  - SD= sqrt(np(1-p)) = sqrt(50*.45*.55) = 3.518
  - Then Z=(30-22.5)/3.518 = 2.132, and we find P(Z>2.132)

```
. di 1-normal(2.132)
.01650342
```

  - Using the continuity correction (subtracting .5 from X)

```
. di 1-normal( (29.5-22.5)/3.518)
.02330831
```

Normal approximation to the Binomial distribution

Binomial distribution, now=100

- Using the binomial distribution for P(X>=60; n=100, p=.45)

```
. di binomialtail(100,60,.45)
.00182018
```

Normal approximation to the Binomial distribution, now=100

- Using the Normal approximation
  - Mean=100*.45=45
  - SD= sqrt(100*.45*.55) = 4.975
  - Then Z=(60-45)/4.975 = 3.015, and we find P(Z>3.015)

```
. di 1-normal(3.015)
.0012849
```

  - Using the continuity correction (subtracting .5 from X)

```
. di 1-normal( (59.5-45)/4.975)
.00178088
```

Sampling distribution of a proportion

- Previous slides were about estimating X, the number of successes
- We often are more interested in the *proportion p* of successes, rather than the *number* of successes x
- The true population proportion p is estimated by

$$\hat{p} = x / n$$

x = the number of successes or events

n=the number of trials or people or observations

Sampling distribution of a proportion

- If we take repeated samples of size n from a variable that follows the Bernoulli distribution (i.e. the outcome is 0 or 1), and calculate $\hat{p}=x/n$ for each of the samples (x=total count of successes), if n is large enough, then $\hat{p}$ will follow a normal distribution (by the central limit theorem)

  – The mean of this distribution is p
  – The standard deviation is $\sqrt{\frac{p(1-p)}{n}}$ which is also called the standard error

Reminder: Sampling distribution of a mean

- If we take repeated samples of size n, and calculate $\bar{X}$ for each of the samples, if n is large enough, the $\bar{x}$ s will follow a normal distribution (by the central limit theorem)
  - The mean of this distribution is μ
  - The standard deviation is $\sigma/\sqrt{n}$, which is also called the standard error

Sampling distribution of proportions

- So if p̂ follows a normal distribution with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$
- Then $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ ~ N(0,1)
- This holds true by the CLT
- Considered valid when the expected number of successes np is ≥5 and the expected number of failures n(1-p) is ≥5

Sampling distribution of proportions

So now we can use the normal distribution to calculate probabilities of observing certain <u>proportions</u> in a sample

- E.g. What proportion of samples of size 50 from a population with p=.10 will have a p̂ of .20 or higher?

- What is $P(\hat{p} \geq 0.20)$?
  - Mean=0.10
  - $SE = \sqrt{(.10*.90)}/\sqrt{50} = 0.0424$
  - $P(Z \geq ((.20-.10)/.0424)) = P(Z \geq 2.36)$

```
. display 1-normal(2.36)
.00913747
```

Confidence intervals for proportions

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \text{N(0,1)}$$

– So $$P(-1.96 \le \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le 1.96) = 0.95$$

- Rearranging, we get

$$P(\hat{p} - 1.96\sqrt{p(1-p)/n} \le p \le \hat{p} + 1.96\sqrt{p(1-p)/n}) = 0.95$$

- →
- Lower 95% confidence limit: $\hat{p} - 1.96 * \sqrt{\frac{p(1-p)}{n}}$
- Upper 95% confidence limit: $\hat{p} + 1.96 * \sqrt{\frac{p(1-p)}{n}}$

Confidence intervals for proportions

- However we don't know p (if we did we wouldn't be calculating these intervals). So we substitute p̂ into the formula for the SEM.

- Lower 95% confidence limit: $$\hat{p} - 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Upper 95% confidence limit: $$\hat{p} + 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- With repeated sampling, 95% of such intervals would include the true population parameter p

Confidence intervals for proportions

- Proportion of Biostat 200-2016 students who are parents
  - x with kids = 8 of 55 who answered the survey
  - Proportion with any children = 8/55 = .146
  - Standard error estimate = sqrt [.146*(1-.146)/55] = 0.048
  - 95% CI : (.146 – 1.96*.048, .146+ 1.96*.048) = (.052, .240)

Confidence intervals for proportions in Stata

CI using the normal approximation

```
ci means children_any

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
children_any |         55     .1454545    .0479771        .0492662    .2416429
```

CI proportion

```
ci proportion children_any

                                                          -- Binomial Exact --
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
children_any |         55     .1454545     .047539        .0649514    .2666323
```

Key points

- It is not practical or feasible to study an entire population, so we take a sample
- We need to make inference (i.e. decisions) from our sample to the population
- We use the properties of repeated samples to do so
- For any random variable X with mean μ and standard deviation σ, if the sample n is large enough, the distribution of the sample mean is normally distributed with mean μ and standard deviation $\sigma / \sqrt{n}$
- We use this to calculate intervals with known probability of containing the population mean or proportion

Next Time

- Review Confidence Intervals and sampling distributions
- Hypothesis testing
- Types of error