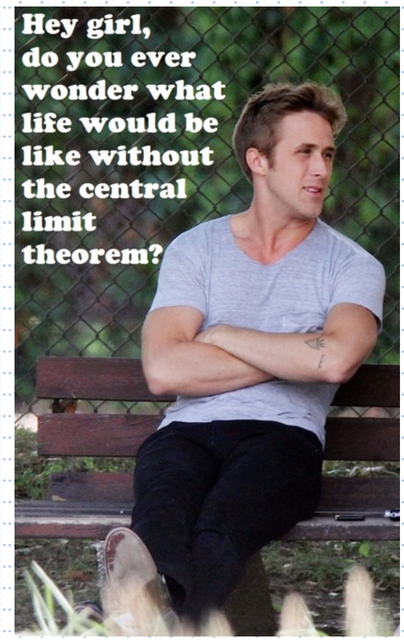
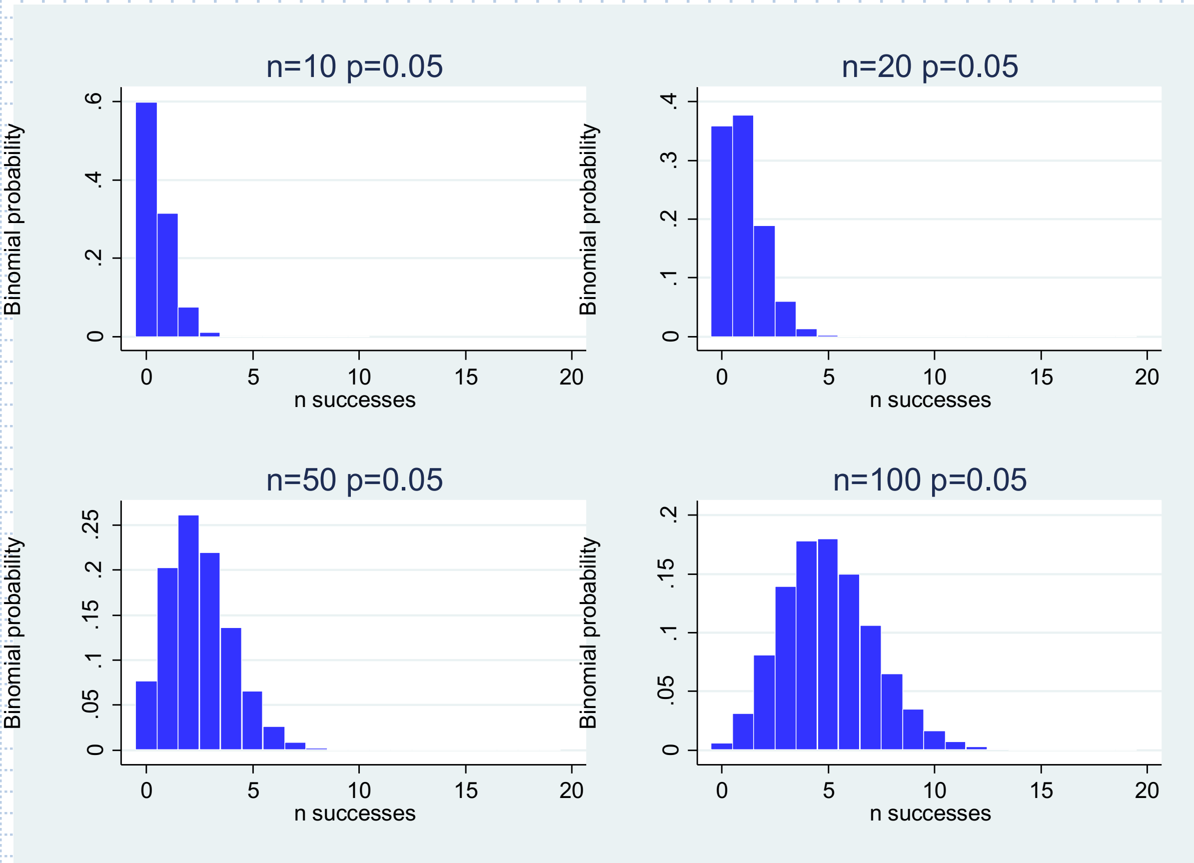
Confidence Intervals
Normal approximation to the Binomial



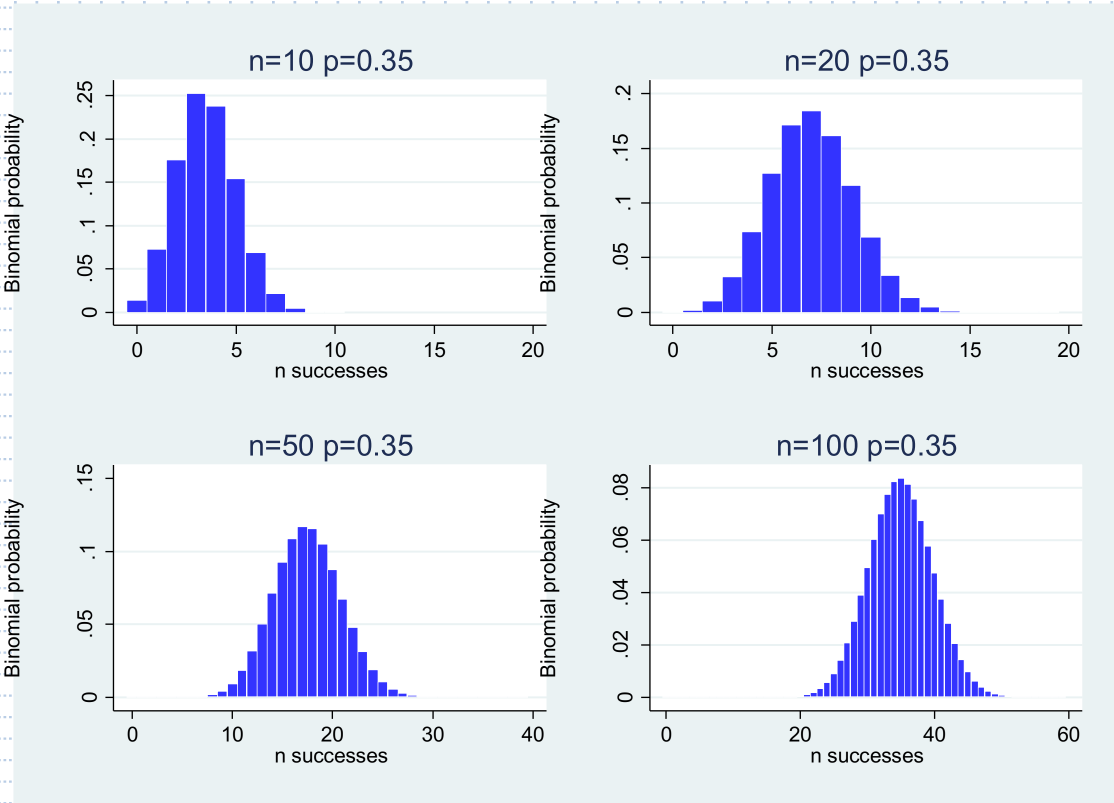
Normal approximation to the Binomial distribution

- 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=$\sqrt{np(1-p)}$
- 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

- 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
```

Now with n=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, n=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
```

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

Sampling distribution of proportions

- So if p̂ follows a normal distribution with mean p and standard deviation
  $$\sqrt{\frac{p(1-p)}{n}}$$
- Then $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ ~ N(0,1) (from 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 $\hat{p}$ 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

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

- Rearranging, we get $P(-1.96 \le \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le 1.96) = 0.95$

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

- 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

- Proportion of Biostat 200-2019 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)

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 on Confidence Intervals

- 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

- Confidence Intervals & sampling distributions
- Hypothesis testing
- Types of error: calculating sample size & Power