Stat 515: Introduction to Statistics

Chapter 6

Recall Definitions from Ch 2

- **Statistic**: numerical summary of a sample
  - Mean($\bar{x}$), proportion($\hat{p}$),median, mode, standard deviation(s), variance(s^2), Q1, Q3, IQR, etc.
  - We use US alphabet letters to denote these
- **Parameter**: numerical summary of a population
  - Mean(μ_x), proportion(ρ), median, mode, standard deviation(σ), variance(σ^2), Q1, Q3, IQR, etc.
  - We usually don't know these values
  - We use Greek letters to denote these

Sampling Distributions

- A **sampling distribution** is the **probability distribution** that specifies probabilities for the possible values of the sample mean or proportion.

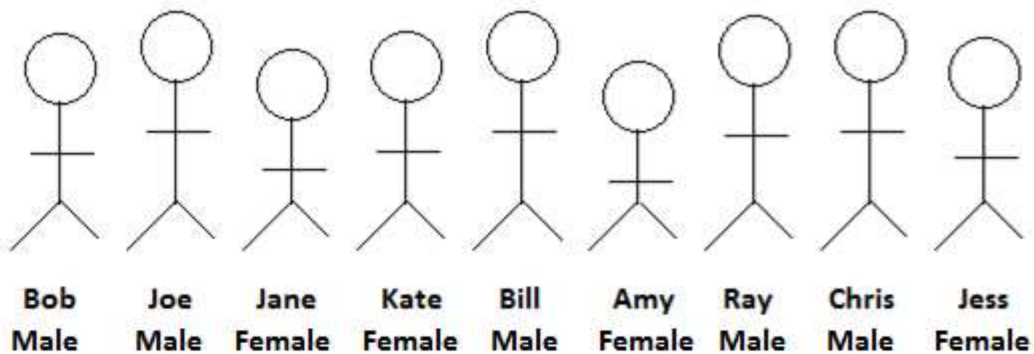
- A **sampling distribution** is a special case of a probability distribution where the outcome of an experiment that we are interested in is a sample statistic such as a **sample proportion($\hat{p}$)** or **sample mean ($\bar{x}$)**
  - It's the same as what we were doing before, but now instead of singular observations we're looking at groups

Sampling Distributions

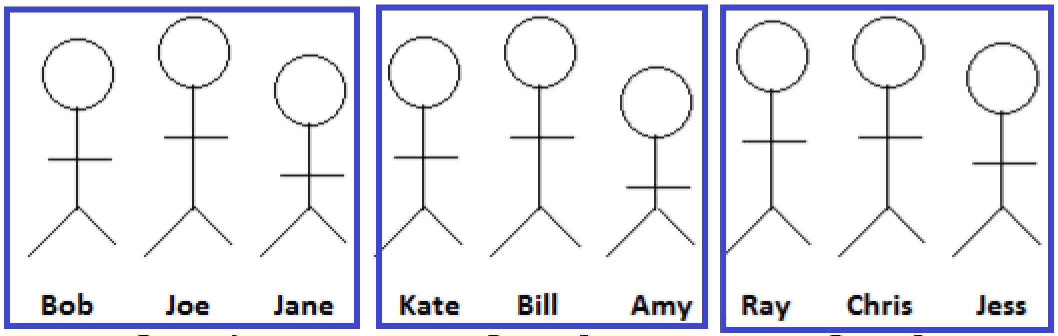
- This is confusing.
  - Remember, before we talked about events and random variables in n trials
  - Now, we're talking about m groups of n trials which yield m sample means or m sample proportions
    - $\overline{x_i} = \frac{\sum x}{n} \; for \; i = 1, 2, \dots, m$
    - $\widehat{p_i} = \frac{x}{n} \; for \; i = 1, 2, \dots, m$

Sampling Distributions

- Variable: Gender of Students
  - Before, we measured individuals:



  - Now, we have one measurement across groups:



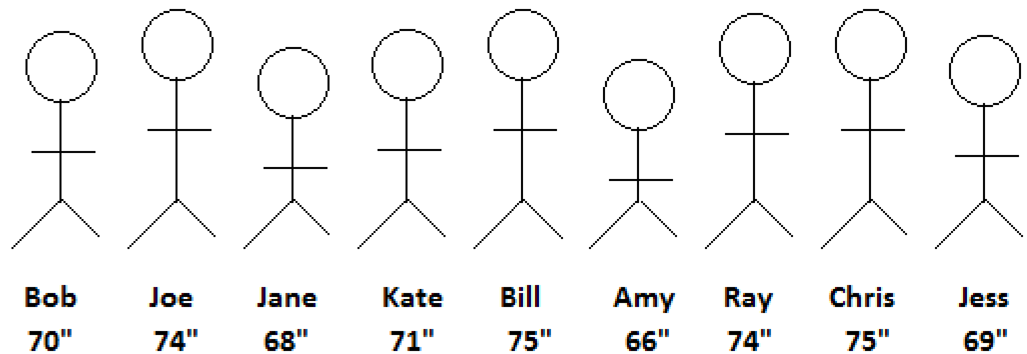
Group 1 $\hat{p}$ = 2/3 = .6667

Group 2 $\hat{p}$ = 1/3 = .3333

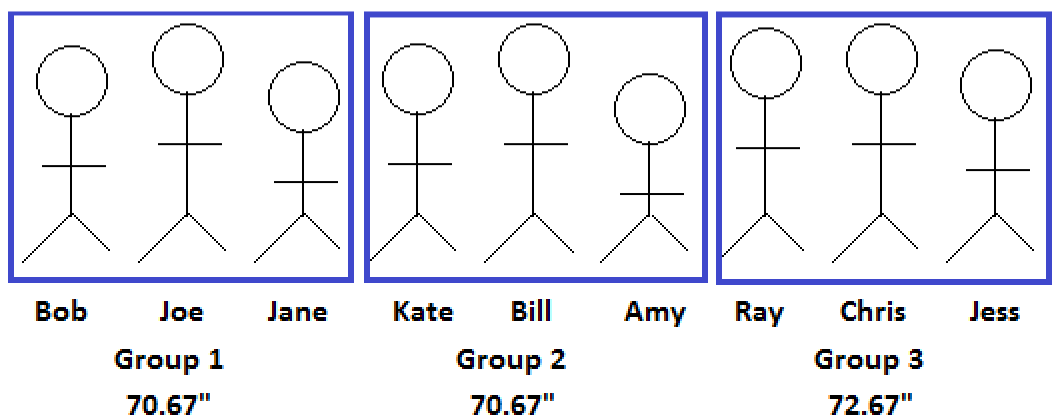
Group3 $\hat{p}$ = 2/3 = .6666

Sampling Distributions

- Variable: Heights of Americans
  - Before, we measured individuals:

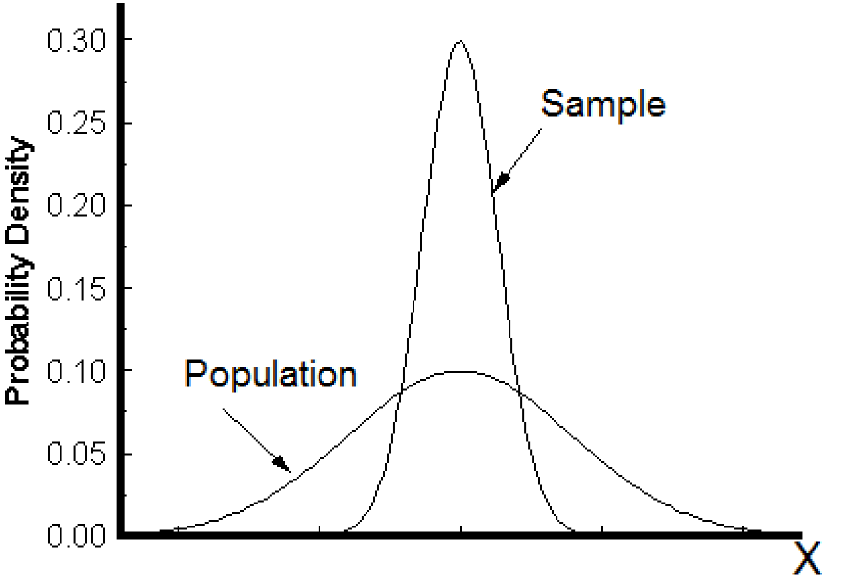


  - Now, we have one measurement across groups:



Sampling Distribution - Graphs

- Sample vs. Population: the sampling distribution is narrower than the population because grouping the data reduces the variation; pay attention to the standard error equations



Sampling Distributions: Proportions

- This first sampling distribution we'll talk about is the **sampling distribution for the sample proportion $\hat{p}$**.

- The idea is that there is some **true population proportion out there, ρ,** but in most cases it isn't feasible to know it
  - We may not have enough time or money to poll the population
  - It may be infeasible to get a population measure

Sampling Distributions: Proportions

- We look at **sample proportions, $\hat{p}$,** the proportion of observations in our sample that have a certain characteristic among our sample
  - Think "x out of n" then $\hat{p} = \frac{x}{n}$

- We've looked at this before in the **descriptive statistics** but now we're going to talk about **all possible sample proportions from repeated random samples from the population** and their distribution (mean and standard deviation)

Sampling Distributions: Proportions

- **Before we had categorical observations:** $x_1, x_2, x_3, \dots, x_n$
  - We would summarize all x's with one **sample proportion, one $\hat{\boldsymbol{p}}$**
  - $\hat{p} = \frac{\text{number of x with desired trait}}{\text{total sample size}}$

    $=$ the proportion of our sample with the desired trait

Sampling Distributions: Proportions

- **Now we have m groups of n subjects with categorical observations:** $\{x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}\}, \{x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}\}, \dots, \{x_{m,1}, x_{m,2}, x_{m,3}, \dots, x_{m,n}\}$
- Now, we find summary statistics for each group

  $\widehat{p_1}, \widehat{p_2}, \widehat{p_3}, \widehat{p_4}, \dots, \widehat{p_m}$

  - We have m sample proportions , one $\hat{p}$ for each group
  - $\widehat{p_1} = \frac{\text{number of x with desired trait in group 1}}{\text{total sample size } of\ group\ 1}$
  - $\widehat{p_2} = \frac{\text{number of x with desired trait in group 2}}{\text{total sample size of } group\ 2}$ …
  - $\widehat{p_m} = \frac{\text{number of x with desired trait in group } m}{\text{total sample size of group m}}$

Sampling Distributions: Proportions

- You could think of each group as a barrel and we're only interested in the proportion of each barrel; we are no longer interested in the individual responses like we might have been before
- The example below shows how we could summarize 40 observations by splitting them into four representative sample proportions

$\widehat{p_1}$	$\widehat{p_2}$	$\widehat{p_3}$	$\widehat{p_4}$
$x_1, x_2, \dots, x_{10}$	$x_{11}, x_{12}, \dots, x_{20}$	$x_{21}, x_{22}, \dots, x_{30}$	$x_{31}, x_{32}, \dots, x_{40}$

Sampling Distribution – Mean and SD

- The **mean of the sampling distribution** for a sample proportion will always equal the population proportion: $\boldsymbol{\mu_{\hat{p}} = \rho}$
  - Even though we know the mean is the population proportion, we note that some $\hat{p}$ will be lower and some will be higher

Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If the population proportion of females in the United States is 51% what would you expect the number of females to be in a random sample of 100 Americans?
  - **A:** 51%, or 51 of 100, is our best guess; think of the binomial expectation.

- Later, we'll do this the other way around and we will call $\hat{p}$ the **point estimate for $\boldsymbol{\rho}$** since it's our best guess for the population proportion if we don't know it

Sampling Distribution – Mean and SD

- The **standard error**, the standard deviation of all possible sample proportions, is:

$$\boldsymbol{\sigma_{\hat{p}}} = \sqrt{\frac{\boldsymbol{\rho(1-\rho)}}{\boldsymbol{n}}}$$

$$= \boldsymbol{St.Dev}(\widehat{p_1}, \widehat{p_2}, \widehat{p_3}, \widehat{p_4}, \ldots, \widehat{p_m})$$

Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If our best guess for $\boldsymbol{\rho}$ is $\hat{p}$ we need a **measure of reliability** for our estimate
  - **A:** We'll talk more about this later, but our standard error calculator is a big part of this

- Recall: $\boldsymbol{\sigma_{\hat{p}}} = \sqrt{\frac{\boldsymbol{\rho(1-\rho)}}{\boldsymbol{n}}}$
- Later, in the case we don't know **ρ** we're estimating it with our **point estimate** $\hat{p}$
  - Consider:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sampling Distribution – Mean and SD

- $\boldsymbol{\mu_{\hat{p}} = \rho}$
  - Even though we know the mean is the population proportion, we note that some $\hat{p}$ will be lower and some will be higher
- $\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{\rho(1-\rho)}{n}}}$
  - Aside:
    - What if we increase n?
      - The standard deviation shrinks
    - What if we decrease n?
      - The standard deviation grows

Sampling Distribution:

- Now that we know the mean and standard deviation of the sample proportions we can calculate z-scores to find some probabilities associated with sample proportions just like we did before.

$$\boldsymbol{\mu_{\hat{p}} = \rho}$$

$$\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{\rho(1-\rho)}{n}}}$$

$$z = \frac{observation - mean}{st.dev} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - \boldsymbol{\rho}}{\sqrt{\frac{\boldsymbol{\rho(1-\rho)}}{\boldsymbol{n}}}}$$

Sampling Distribution:

$$P(\hat{p} > c) = 1 - P\left(z < \frac{c - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = 1 - P\left(z < \frac{c - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

$$P(\hat{p} < c) = P\left(z < \frac{c - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = P\left(z < \frac{c - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

$$P(c_1 < \hat{p} < c_2) = P\left(z < \frac{c_2 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) - P\left(z < \frac{c_1 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right)$$

$$= P\left(z < \frac{c_2 - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right) - P\left(z < \frac{c_1 - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

Sampling Distributions - Example

- Say, we know that **16% of Americans approve of Congress (Gallup).**
- **What is the sampling distribution of the sample proportion** of Americans that approve of Congress for n=100?
  - Note, we aren't interested in the yes or no's individually but the proportion among the ten
  - Here, X=the proportion of the one hundred Americans in each group

Sampling Distributions - Example

- Say, we know that **16% of Americans approve of Congress (Gallup).**
- **What is the sampling distribution of the sample proportion** of Americans that approve of Congress for n=100?

  - n = sample size = **sample size of one hundred**= 100
  - p = population proportion = **16%** =.16

Sampling Distributions – Example

- Let's find the sampling distribution mean:

- **The mean of all sample proportions of n=100** = $\mu_{\hat{p}} = \rho = 16\% = .16$

  – Some $\hat{p}$ will be lower and some will be higher but **the mean of all sample proportions of n=100 will be .6**

Sampling Distributions – Example

- Let's find the sampling distribution st. error:

- **The st. deviation of all sample proportions of n=100**

= Standard Error = $\sigma_{\hat{p}} = \sqrt{\frac{\rho(1-\rho)}{n}}$

$$= \sqrt{\frac{.16(1-.16)}{100}} = .0367$$

Sampling Distributions – Example

- Let's find the sampling distribution :

$$\mu_{\hat{p}} = \rho = 16\% = .16$$

$$\sigma_{\hat{p}} = \sqrt{\frac{\rho(1-\rho)}{n}} = \sqrt{\frac{.16(1-.16)}{100}} = .0367$$

Sampling Distributions – Example

- The probability that **most**, of our sample of n=100, approve of Congress:

R

$$
\begin{aligned}
P(\hat{p} > .5) &= 1 - \mathrm{P}(\hat{p} < .5) \\
&= 1 - pnorm(.5, .16, .0367\ ) \\
&\approx 0
\end{aligned}
$$

Z-table:

$$
\begin{aligned}
P(\hat{p} > .5) &= P\left(z > \frac{.5 - .16}{.0367}\right) = P(Z > 9.26) \\
&= 1 - P(Z \leq 9.26) \approx 1 - 1 \\
&= 0
\end{aligned}
$$

Note: the z-table is a less accurate approximation than R

Sampling Distributions – Example

- The probability that **less than 10%**, of our sample of n=100, approve of Congress:

R

$$P(\hat{p} < .1) = pnorm(.1, .16, .0367)$$
$$= .0510$$

Z-table:

$$P(\hat{p} < .1) = P\left(z < \frac{.1 - .16}{.0367}\right) = P(Z < -1.63)$$
$$= .0516$$

Note: the z-table is a less accurate approximation than R

Sampling Distributions – Example

- The probability that **between 5 and 19 percent**, of our sample of n=100, approve of Congress:

R

$$P(.05 < \hat{p} < .19) = \mathrm{P}(\hat{p} < .19) - \mathrm{P}(\hat{p} < .05)$$
$$= pnorm(.19, .16, .0367) - pnorm(.05, .16, .0367)$$
$$= .7917991$$

Z-table:

$$P(.05 < \hat{p} < .19) = \mathrm{P}(\hat{p} < .19) - \mathrm{P}(\hat{p} < .05)$$
$$= \mathrm{P}\left(\mathrm{z} < \frac{.19 - .16}{.0367}\right) - \mathrm{P}\left(z < \frac{.05 - .16}{.0367}\right)$$
$$= P(Z < .82) - P(Z < -3.00)$$
$$= .7939 - .0013$$
$$= .7926$$

Sampling Distributions – Example

- **Note:** we had to assume normality of $\hat{p}$ to use the Z-score transformation and R code I provided to solve the previous probabilities

- Later, we will see how we are able to make that assumption – unlocking all of the nice methodologies of the Normal distribution

Sampling Distributions: Means

- This second sampling distribution we'll talk about is the **sampling distribution for the sample mean**.

- The idea is that there is some **true population mean out there, μ,** but it might not be feasible to know it
  - We may not have enough time or money to poll the population
  - It may be infeasible to get a population measure

Sampling Distributions: Means

- Instead, we look at **sample mean, $\bar{x}$,** the mean of quantitative observations

- We've looked at this before in the **descriptive statistics** but now we're going to talk about **all possible sample means from repeated random samples from our population**

Sampling Distributions: Means

- **Before we had quantitative observations:** $x_1, x_2, x_3, \ldots, x_n$
  - We would summarize all x's with one **sample mean, one $\boldsymbol{\bar{x}}$**
  - $\bar{x} = \frac{\text{the sum of x's}}{\text{the total sample size}} = \frac{\sum x}{\text{n}}$

    $=$ the mean of the observations in our sample

Sampling Distributions: Means

- **Now we have m groups of n subjects with categorical observations:** $\{x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}\}$,$\{x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}\}$, …,$\{x_{m,1}, x_{m,2}, x_{m,3}, \dots, x_{m,n}\}$
- Now, we find summary statistics for each group

  $\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, \dots, \overline{x_m}$

  – We have m sample means, one $\bar{x}$ for each group

    - $\overline{x_1} = \frac{\text{the sum of x's } from\ group\ 1}{\text{the total sample size of } group\ 1} = \frac{\sum x}{\text{n}}$
    - $\overline{x_2} = \frac{\text{the sum of x's } from\ group\ 2}{\text{the total sample size of group 2}} = \frac{\sum x}{\text{n}} \dots$
    - $\overline{x_m} = \frac{\text{the sum of x's } from\ group\ m}{\text{the total sample size } of\ group\ m} = \frac{\sum x}{\text{n}}$

Sampling Distributions: Means

- You could think of each group as a barrel and we're only interested in the mean of each barrel; we are no longer interested in the individual responses
- The example below shows how we could summarize 40 observations, into four representative sample means

$\overline{x_1}$	$\overline{x_2}$	$\overline{x_3}$	$\overline{x_4}$
$x_1, x_2, \dots, x_{10}$	$x_{11}, x_{12}, \dots, x_{20}$	$x_{21}, x_{22}, \dots, x_{30}$	$x_{31}, x_{32}, \dots, x_{40}$

Sampling Distribution – Mean and SD

- The mean of the sampling distribution for a sample mean will always equal the population mean: $\boldsymbol{\mu_{\bar{x}} = \mu_x}$
  - This is the mean of all possible sample means, but we note that some $\boldsymbol{\bar{x}}$ will be lower and some will be higher

Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If the population mean of time Americans spend on social media is 100 minutes with a standard deviation of 25 minutes what would you expect the average time a sample of 35 Americans spent on social media?
  - **A:** 100 minutes is our best guess.

- Later, we'll do this the other way around and we will call $\bar{x}$ the **point estimate for $\boldsymbol{\mu_x}$** since it's our best guess for the population mean if we don't know it

Sampling Distribution – Mean and SD

- The standard error, the standard deviation of all possible sample means, is:

$$\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

$$\boldsymbol{= St.Dev}(\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, \ldots, \overline{x_m})$$

Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If our best guess for $\boldsymbol{\mu}$ is $\bar{x}$ we need a **measure of reliability** for our estimate
  - **A:** We'll talk more about this later, but our standard error calculator is a big part of this

- Later, in the case we don't know μ_x or σ_x we're estimating it with our **point estimate** $\boldsymbol{\bar{x}}$
  - Recall: $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$
  - Consider: $\boldsymbol{\frac{s_x}{\sqrt{n}}}$ **[Note: we estimate $\boldsymbol{\sigma_x = s_x}$]**

Sampling Distribution – Mean and SD

- $\boldsymbol{\mu_{\overline{x}} = mean\ of\ all\ sample\ means = \mu_x}$
  - Even though we know the mean is the population mean, we note that some $\bar{x}$ will be lower and some will be higher
- $\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}}$
- Aside:
  - What if we increase n?
    - The standard deviation shrinks
  - What if we decrease n?
    - The standard deviation grows

Sampling Distribution:

- Now that we know the mean and standard error of the sample means we can calculate z-scores to find some probabilities associated with sample means just like we did before.

$$\boldsymbol{\mu_{\bar{x}} = \mu_x}$$

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

$$z = \frac{observation - mean}{st.dev} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

Sampling Distribution:

$$P(\bar{x} > c) = 1 - P\left(z < \frac{c - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = 1 - P\left(z < \frac{c - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

$$P(\bar{x} < c) = P\left(z < \frac{c - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(z < \frac{c - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

$$P(c_1 < \bar{x} < c_2) = P\left(z < \frac{c_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) - P\left(z < \frac{c_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P\left(z < \frac{c_2 - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right) - P\left(z < \frac{c_1 - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

Sampling Distributions - Example

- Say, we know that **the average American spends 100 minutes on social media per day with a standard deviation of 25 minutes.**
- **What is the sampling distribution of the sample mean** of time Americans spend on social media for n=35?
  - Note, we aren't interested in the individuals but the group of thirty five
  - Here, X=the proportion of the ten Americans in each group

Sampling Distributions - Example

- Say, we know that **the average American spends 100 minutes on social media per day with a standard deviation of 25 minutes.**
- **What is the sampling distribution of the sample mean** of time Americans spend on social media for n=35?

  - n = sample size = **sample size of thirty five** = 35
  - μ_x = population mean = 100
  - σ_x = population standard deviation = 25

Sampling Distributions – Example

- <u>Let's find the sampling distribution mean:</u>

- **The mean of all sample ,eams of n=35**
  = $\mu_{\bar{x}} = \mu_x = 100$

  – Some $\boldsymbol{\bar{x}}$ will be lower and some will be higher but **the mean of all sample means of n=35 will be 100**

Sampling Distributions – Example

- Let's find the sampling distribution st. error:

- **The st. deviation of all sample means of n=35**
  = Standard Error

$$= \boldsymbol{\sigma_{\bar{x}}} = \frac{\boldsymbol{\sigma_x}}{\sqrt{\boldsymbol{n}}} = \frac{25}{\sqrt{35}} = 4.2258$$

Sampling Distributions – Example

- <u>Let's find the sampling distribution :</u>

$$\mu_{\bar{x}} = \mu_x = 100$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{25}{\sqrt{35}} = 4.2258$$

Sampling Distributions – Example

- The probability that a sample of n=35 spend **more than two hours** on social media on average:

<u>R</u>

$$P(\bar{x} > 120) = 1 - \mathrm{P}(\bar{x} < 120)$$
$$= 1 - pnorm(120{,}100{,}4.2258)$$
$$= .000001107046$$

<u>Z-table:</u>

$$P(\bar{x} > 120) = P\left(z > \frac{120 - 100}{4.2258}\right) = P(Z > 4.73)$$
$$= 1 - P(Z < 4.73) \approx 1 - 1$$
$$= 0$$

Note: the z-table is a less accurate approximation than R

Sampling Distributions – Example

- The probability that a sample of n=35 spend **less than one hour** on social media on average:

<u>R</u>

$$P(\bar{x} < 60) = \mathrm{P}(\bar{x} < 60)$$
$$= pnorm(60{,}100{,}4.2258)$$
$$= 1.458519 * 10^{-21}$$

<u>Z-table:</u>

$$P(\bar{x} > 60) = P\left(z < \frac{60 - 100}{4.2258}\right) = P(Z < -9.47)$$
$$= P(Z < -9.47)$$
$$\approx 0$$

Note: the z-table is a less accurate approximation than R

Sampling Distributions – Example

- The probability that a sample of n=35 spend **between 1 and 1.5 hours** on social media on average:

<u>R</u>

$$P(60 < \bar{x} < 90) = \mathrm{P}(\bar{x} < 90) - \mathrm{P}(\bar{x} < 60)$$
$$= pnorm(90{,}100{,}4.2258) - pnorm(60{,}100{,}4.2258)$$
$$= .008980629$$

<u>Z-table:</u>

$$P(60 < \bar{x} < 90) = P\left(\frac{90-100}{4.2258} < z < \frac{60-100}{4.2258}\right)$$
$$= P(Z < -2.37) - P(Z < -9.47)$$
$$\approx .0089 - 0$$
$$= 0$$

Note: the z-table is a less accurate approximation than R

Sampling Distributions – Example

- **Note:** we had to assume normality of $\bar{x}$ to use the Z-score transformation and R code I provided to solve the previous probabilities

- Later, we will see how we are able to make that assumption – unlocking all of the nice methodologies of the Normal distribution

Recall: Law of Large Numbers

- (**LLN 1**) – As the sample size increases the sample estimates ($\bar{x}$ or $\hat{p}$) approach the population values (μ or σ)

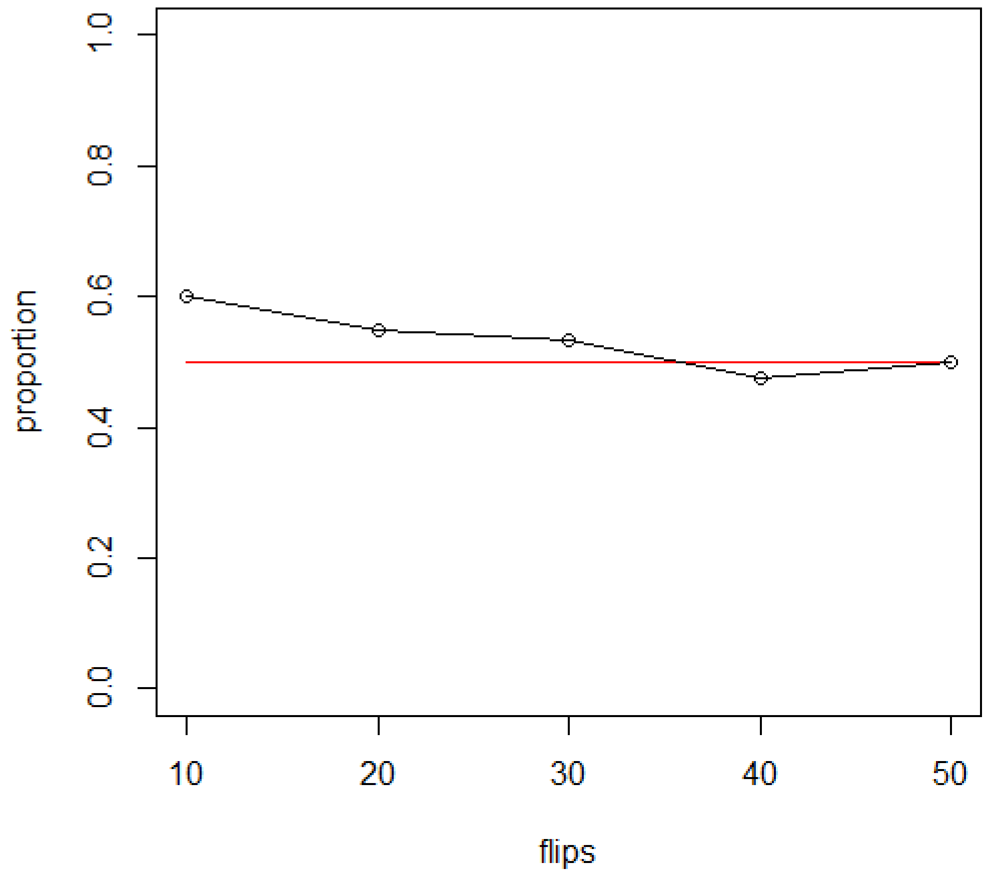
- (**LLN 2**) – As the number of trials increase the **proportion** of occurrences of any given outcome approaches the probability in the long run.

Recall: Law of Large Numbers

- 10 flips: 6 heads were flipped
  - $Total\ proportion\ = \frac{x}{n} = \frac{6}{10} = .60 = 60\%\ heads$
- 10 more flips: 5 heads were flipped
  - $Total\ proportion\ = \frac{x}{n} = \frac{5+6}{10+10} = \frac{11}{20} = .55 = 55\%\ heads$
- 10 more flips: 5 heads were flipped
  - $Total\ proportion\ = \frac{x}{n} = \frac{11+5}{20+10} = \frac{16}{30} = .5333 = 53.33\%\ heads$
- 10 more flips: 3 heads were flipped
  - $Total\ proportion\ = \frac{x}{n} = \frac{16+3}{30+10} = \frac{19}{40} = .475 = 47.5\%\ heads$
- 10 more flips: 6 heads were flipped
  - $Total\ proportion\ = \frac{x}{n} = \frac{19+6}{40+10} = \frac{25}{50} = .5 = 50\%\ heads$

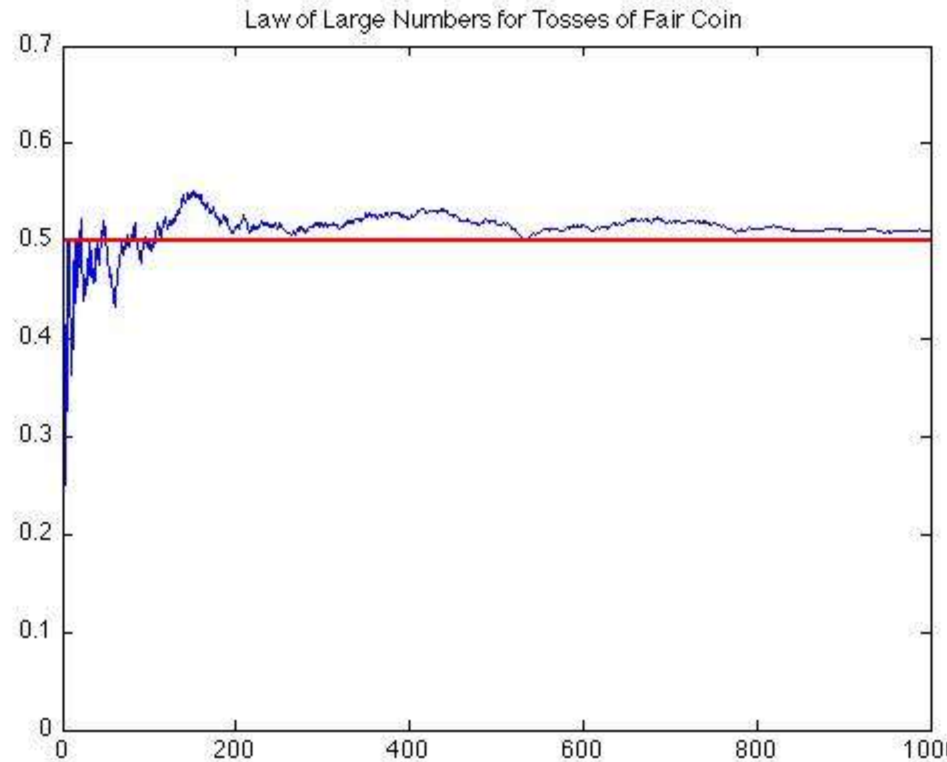
Recall: Law of Large Numbers

- (**LLN**) – As the number of flips increase the **proportion** of heads approaches the **probability** of seeing a heads, P(heads)=.5, which is the red line.



Recall: Law of Large Numbers

- At first the proportion is all over the place – you can see the large spikes in the graph
- Importantly, we see that the proportion of coins that landed on heads levels off and gets closer and closer to 50%, the probability, which is where we expect it to go 'in the long run!'



Central Limit Theorem: Proportions

- For random sampling with a **large sample size n, the sampling distribution of the sample proportion** is approximately a normal distribution
  - $n * p \geq 15 \; and \; n * (1 - p) \geq 15$

- Introduction:
  - https://www.youtube.com/watch?v=Pujol1yC1_A

Central Limit Theorem: Means

- For random sampling with a **large sample size n, the sampling distribution of the sample mean** is approximately a normal distribution
  - For us, 30 is close enough to infinity
- Introduction:
  - https://www.youtube.com/watch?v=Pujol1yC1_A

Central Limit Theorem: Means

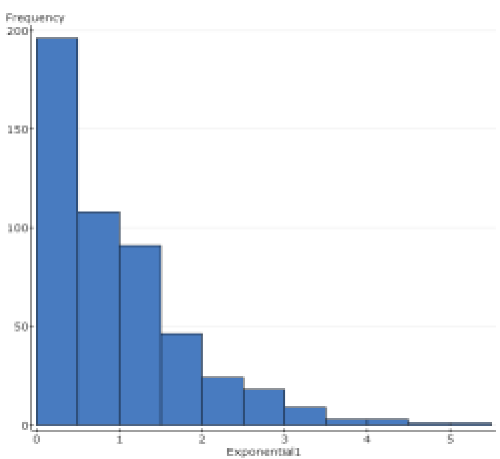
1) For any population the sampling distribution of $\bar{x}$ is bell shaped when the sample size n is large, when n is thirty or more
2) The sampling distribution of $\bar{x}$ is bell-shaped when the population distribution is distribution is bell-shaped, regardless of sample size
3) We do not know the shape of the sampling distribution of $\bar{x}$ if the sample size is small and the population distribution isn't bell-shaped

Central Limit Theorem

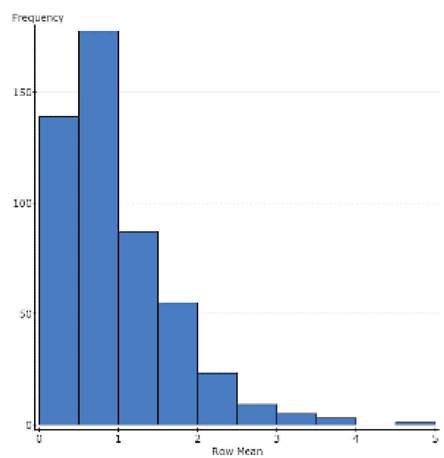
For any population the sampling distribution of $\bar{x}$ is bell shaped when the sample size n is large, when n is thirty or more

Note: for small sample size we can't say this.

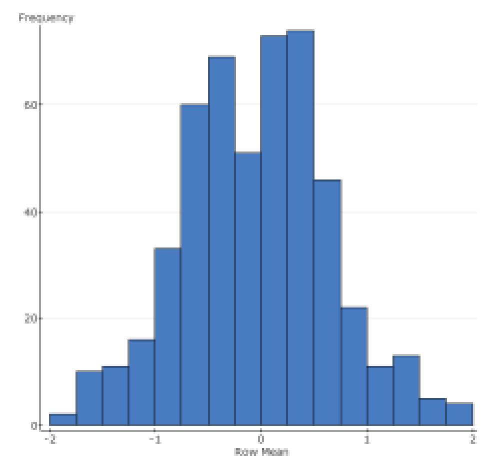
Population



$\bar{x}$ when n=2



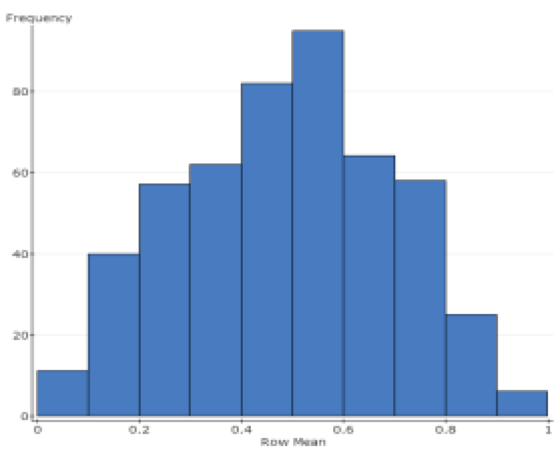
$\bar{x}$ when n=30



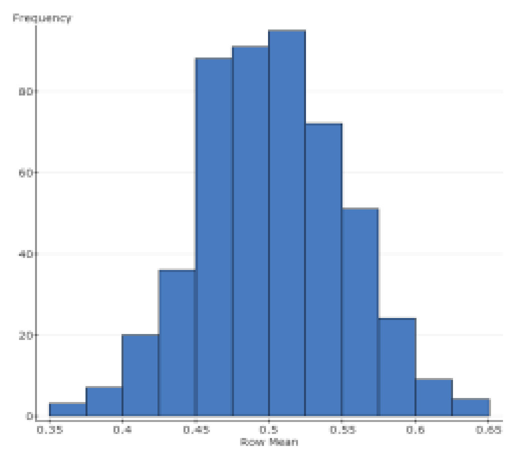
Central Limit Theorem

The sampling distribution of x_{bar} is bell-shaped when the population distribution is distribution is bell-shaped, regardless of sample size

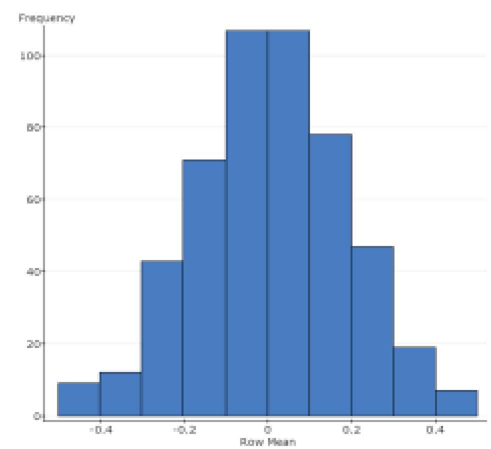
Population



$\bar{x}$ when n=2



$\bar{x}$ when n=30



Chapter 6 Supplement
(not in text book)

Central Limit Theorem

- **Recall:** The sampling distribution of x_{bar} is bell-shaped even for n=1 if X follows the normal distribution
- If a simple random sample is drawn from the population then $z = \frac{\bar{x} - \mu_x}{\sigma_{\overline{x}}}$ follows the standard normal distribution
- The difficulty in this is that we rarely know the value of the parameters: μ_x or σ_x in $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

Central Limit Theorem

- To combat this difficulty we will cover a few more named distributions and sampling distributions
  1. s^2 and the χ^2 (chi-squared) distribution
  2. $\bar{x}, s^2$ and the Student's t-distribution
  3. Two variances and the F distribution

s^2 and the χ^2 (chi-squared) distribution

- The chi-squared density can be defined as follows:

$$f_{x(x)} = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} I(X > 0)$$

- Mean = n
- Variance = $2n$

χ^2 (chi-squared) distribution in R

- $P(X = x) = 0$ as the probability of any one value is always zero

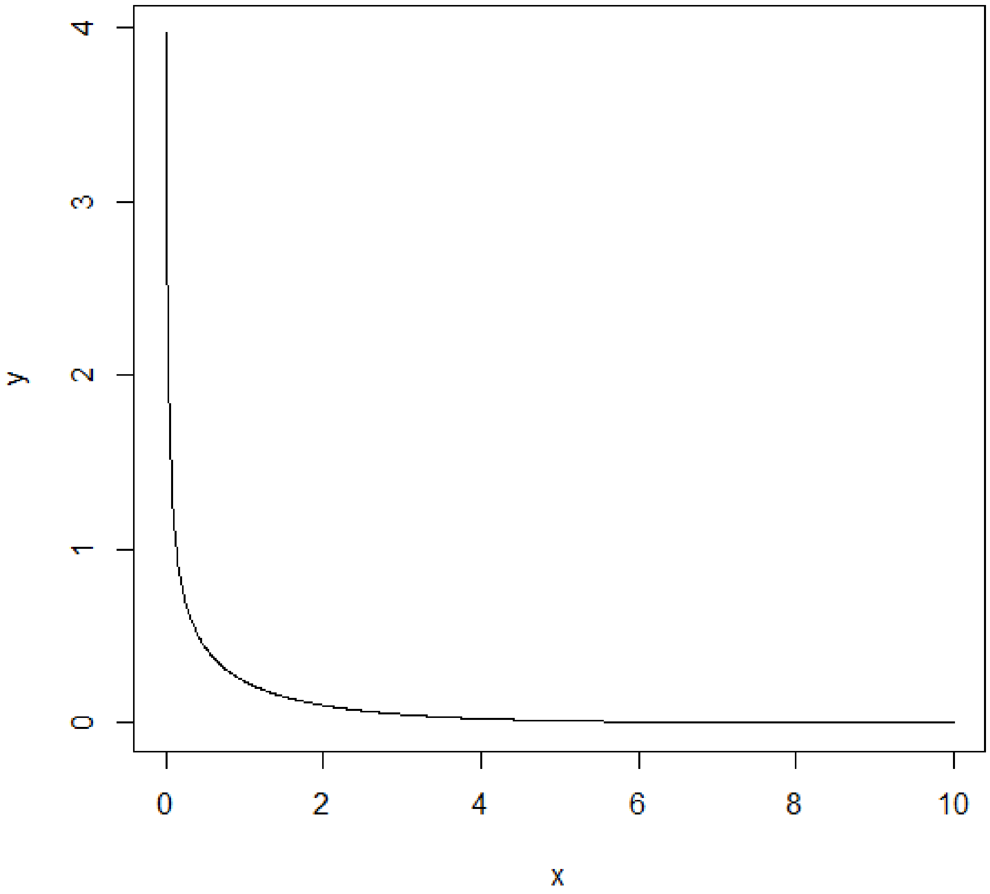
- $P(X \leq x) =$ pchisq(x, n)

- $P(X \geq x) =$ 1 - pchisq(x, n)

- $P(x_1 < X < x_2) =$ pchisq(x_2, n) - pchisq(x_1, n)

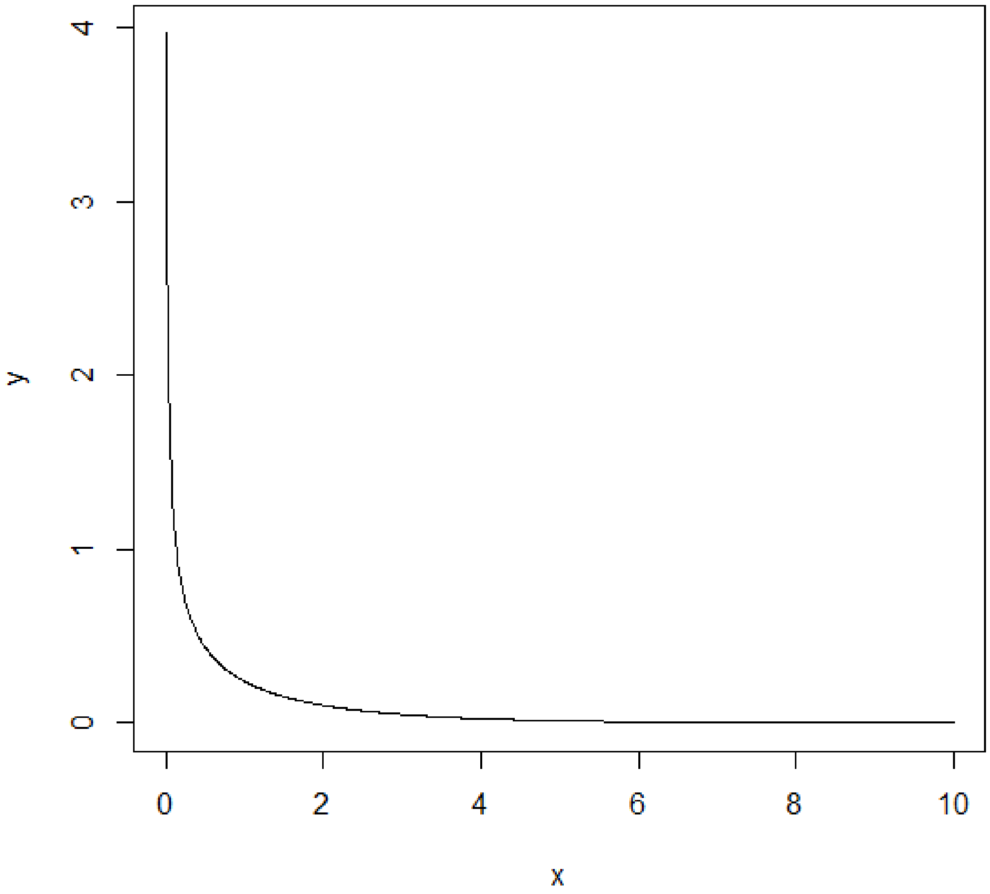
s^2 and the χ^2 (chi-squared) distribution

n=1



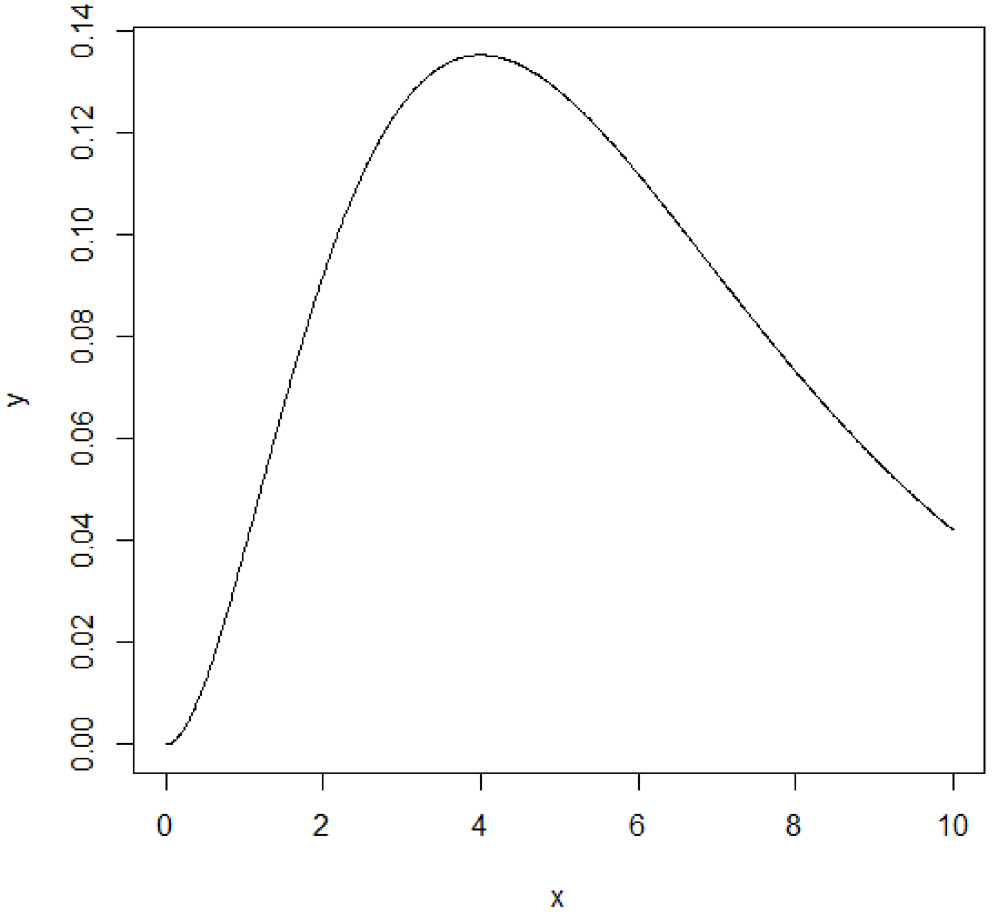
s^2 and the χ^2 (chi-squared) distribution

n=1



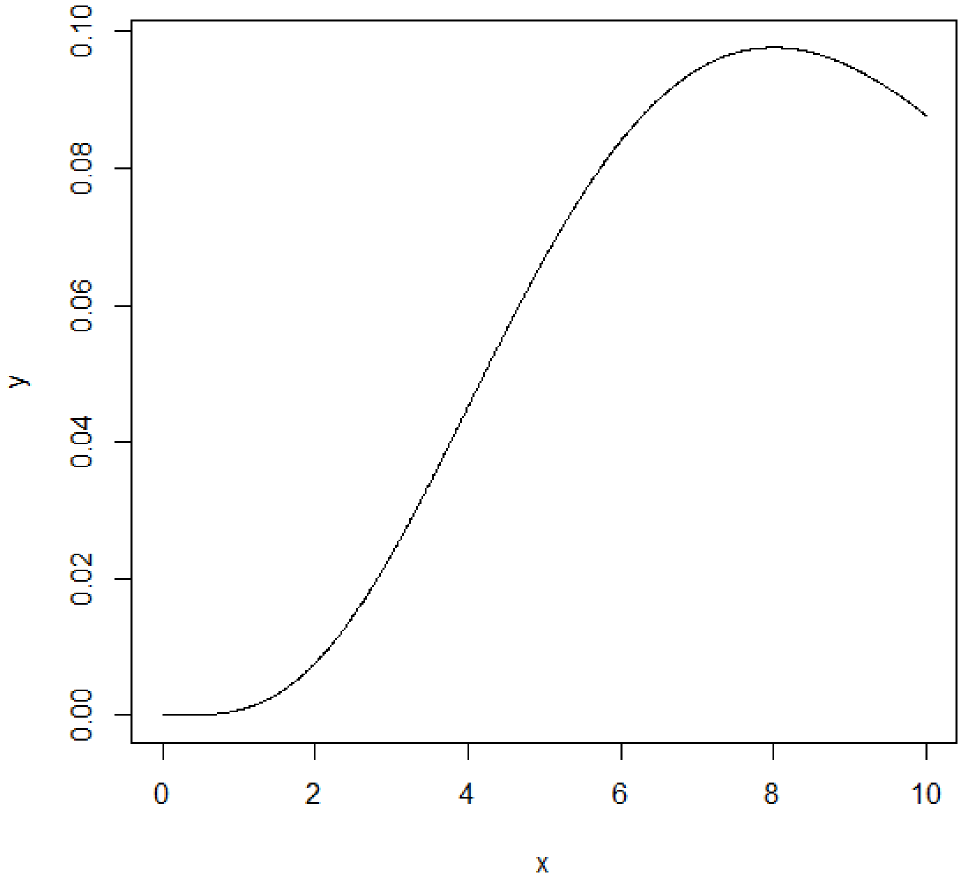
s^2 and the χ^2 (chi-squared) distribution

n=6



s^2 and the χ^2 (chi-squared) distribution

n=10



s^2 and the χ^2 (chi-squared) distribution

- The chi-squared distribution can be defined as follows:

$$\mathrm{X}^2 = z_1^2 + z_2^2 + \cdots + z_n^2$$

- **$\boldsymbol{X^2}$ follows the chi-squared distribution with n degrees of freedom** where z_i are independent random variables that follow the standard normal distribution

s^2 and the χ^2 (chi-squared) distribution

- Consider $x_1, x_2, \dots x_n$ independent from a normal distribution
- If we look at the formula for the sample variance from chapter two we see that we are summing $(x_i - \bar{x})^2$ (n normal random variables)

s^2 and the χ^2 (chi-squared) distribution

- We're close to showing s^2 follows the χ^2 distribution but $(x_i - \bar{x})^2$ are normal – not the standard normal.

- Now, if we can write it as a sum of squared standard normal random variables we can say that s^2 follows the χ^2 distribution
  - This requires a "cute" trick

s^2 and the χ^2 (chi-squared) distribution

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

$$\left(\boldsymbol{\frac{n-1}{\sigma_x^2}}\right) s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}\left(\boldsymbol{\frac{n-1}{\sigma_x^2}}\right)$$

$$\left(\frac{(n-1)s^2}{\sigma_x^2}\right) = \sum\left(\boldsymbol{\frac{x_i - \bar{x}}{\sigma_x}}\right)^2$$

s^2 and the χ^2 (chi-squared) distribution

$$\mathrm{X}_{n-1}^2 = \left(\frac{(n-1)s^2}{\sigma_x^2}\right) = \sum\left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2$$

- Again, consider $\left(\frac{x_i - \bar{x}}{\sigma_x}\right)$ on the right hand side
- If we change $\bar{x}$ to μ we have the usual $\mathrm{z} = \frac{x_i - \mu_x}{\sigma_x}$

- **Note:** We lose one degree of freedom, going from n to n-1, because we are using $\bar{x}$ instead of μ

$\bar{x}, s^2$ and the Student's t-distribution

- Still dealing with the difficulty of needing to know the population standard deviation for the Central limit theorem we talk about the t-distribution

$\bar{x}, s^2$ and the Student's t-distribution

- The t density can be defined as follows:

$$f_{x(x)} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} I(X \epsilon \mathbb{R})$$

- Mean = 0
- Variance = $\frac{n}{n-2}$

Properties of the t-distribution

1. The t-distribution is different for different degrees of freedom
2. The t-distribution is centered and symmetric at 0
3. The area under the curve is 1 and ½ on either side of 0
4. The probability approaches 0 as we move away from 0
5. The t-distribution has fatter tails than the standard normal
6. As the sample size increases t gets close to z

$\bar{x}, s^2$ and the Student's t-distribution in R

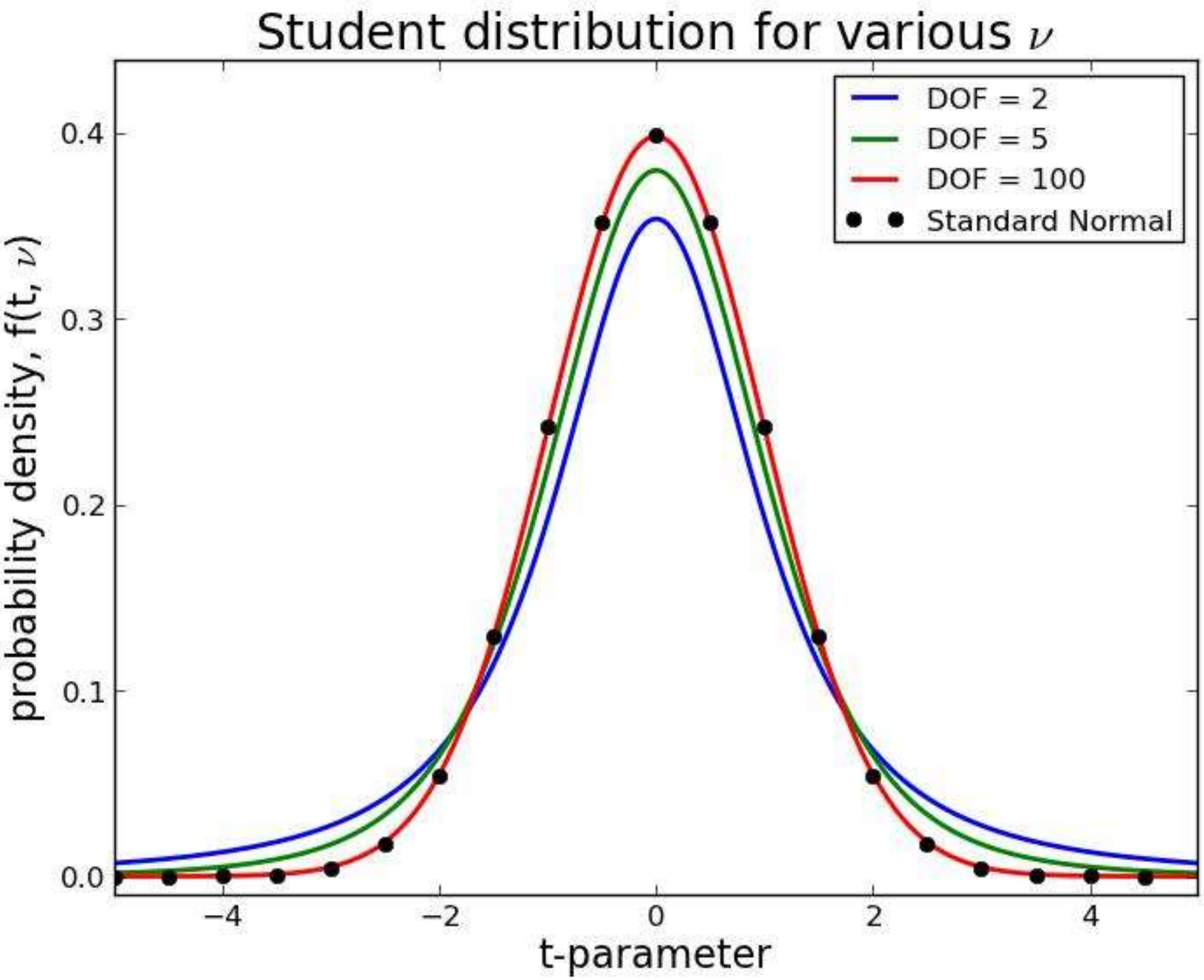
- $P(X = x) = 0$ as the probability of any one value is always zero

- $P(X \leq x) =$ pt(x, n)

- $P(X \geq x) =$ 1 - pt(x, n)

- $P(x_1 < X < x_2) =$ pt(x_2, n) - pt(x_1, n)

The t-distribution



$\bar{x}, {s_x}^2$ and the Student's t-distribution

- The t distribution can be defined as follows:

$$t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

- t_n **follows the t distribution with n degrees of freedom** where Z follows the standard normal distribution and χ^2 is chi-squared and divided by its degrees of freedom

$\bar{x}$, s_x^2 and the Student's t-distribution

- Relating it back to ${s_x}^2$ and χ_n^2

$$t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}} = \frac{\left(\frac{\frac{\bar{x}-\mu_x}{\sigma_x}}{\sqrt{n}}\right)}{\sqrt{\frac{\left(\frac{(n-1){s_x}^2}{\sigma_x^2}\right)}{n-1}}}$$

$\bar{x}$, s_x^2 and the Student's t-distribution

- Relating it back to s_x^2 and χ_n^2

$$t_{n-1} = \frac{\left(\frac{\bar{x}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)}{\sqrt{\frac{s_x^2}{\sigma_x^2}}} = \frac{\left(\frac{\bar{x}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)}{\frac{s_x}{\sigma_x}} = \frac{\left(\frac{\bar{x}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)\frac{\sigma_x}{s_x}}{\frac{s_x}{\sigma_x}\frac{\sigma_x}{s_x}}$$

$$= \frac{\bar{x}-\mu_x}{(s_x/\sqrt{n})}$$

$\bar{x}$, s_x^2 and the Student's t-distribution

- Note the similarity to the z-score: the only difference here is that we estimate σ_x with s_x
- Later we will use the t-distribution when we make inference on the mean and don't know the population standard deviation

$$t_{n-1} = \frac{\bar{x} - \mu_x}{(s_x/\sqrt{n})}$$

Two variances and the F distribution

- This result is something we will use in Chapter 11 and deals with the relationship between the F and the t distribution.

Two variances and the F distribution

- The F density can be defined as follows:

$$f_{x(x)} = \frac{\left(\sqrt{\frac{(d_1 x)^2 d_2^2}{(d_1 x + d_2)^{(d_1+d_2)}}}\right)}{xB\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} I(x \geq 0)$$

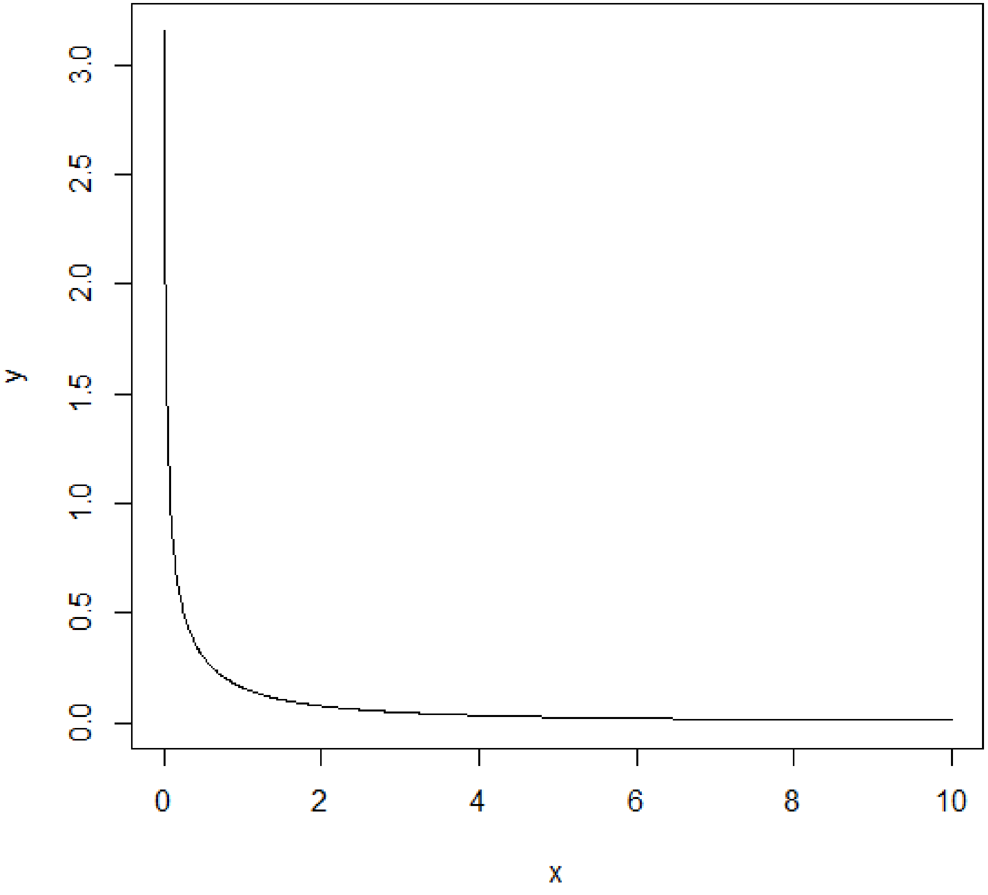
- Mean = $\frac{d_2}{d_2-2}$ for d_2>2
- Variance =$\frac{2 * d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ for d_2 >4

Two variances and the F distribution in R

- $P(X = x) = 0$ as the probability of any one value is always zero
- $P(X \leq x) =$ pf(x, n_x, n_y)
- $P(X \geq x) =$ 1 - pf(x, n_x, n_y)
- $P(x_1 < X < x_2) =$ pf(x_2, n_x, n_y) - pf(x_1, n_x, n_y)

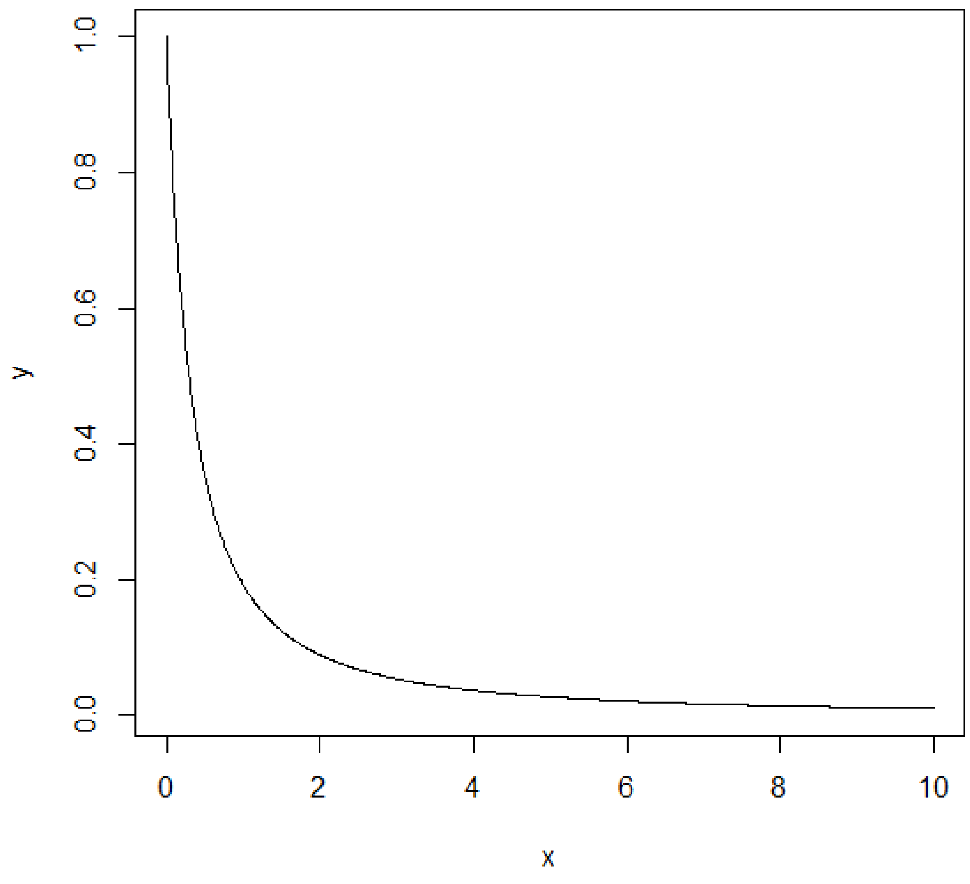
Two variances and the F distribution

$n_x = 1, n_y = 1$



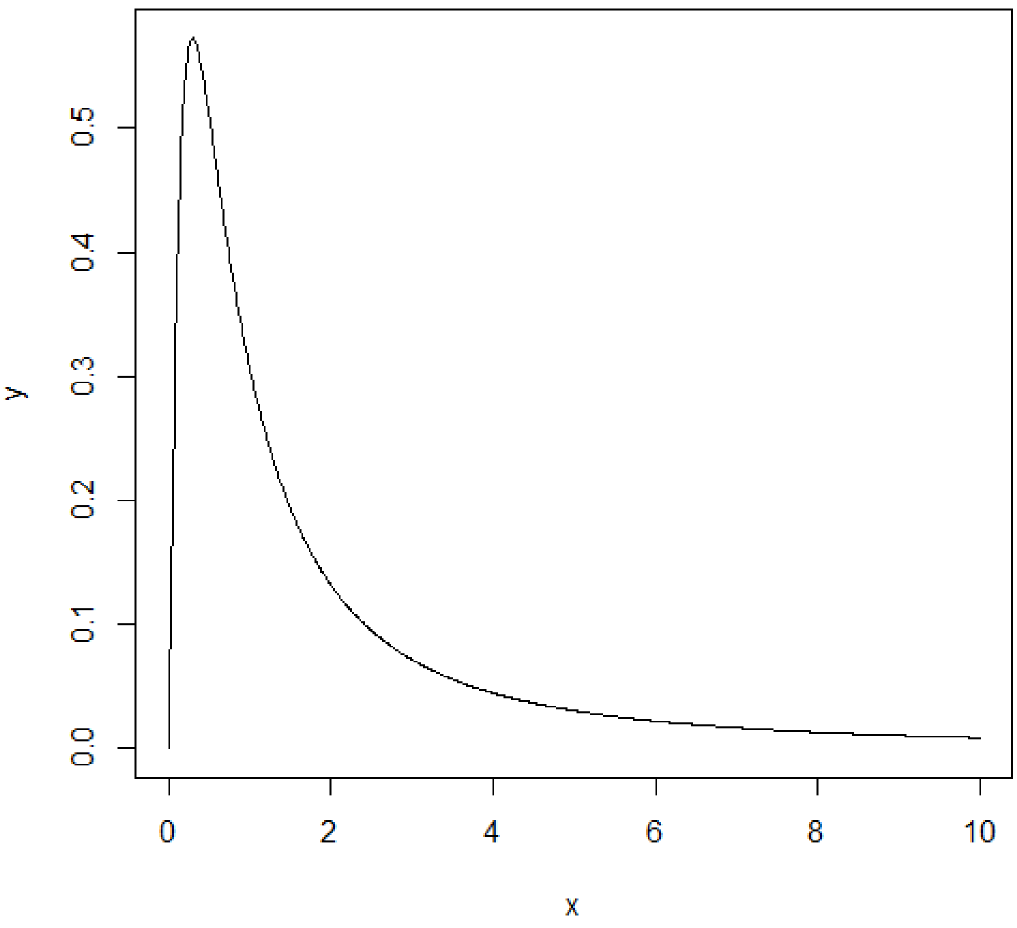
Two variances and the F distribution

$n_x = 2, n_y = 1$



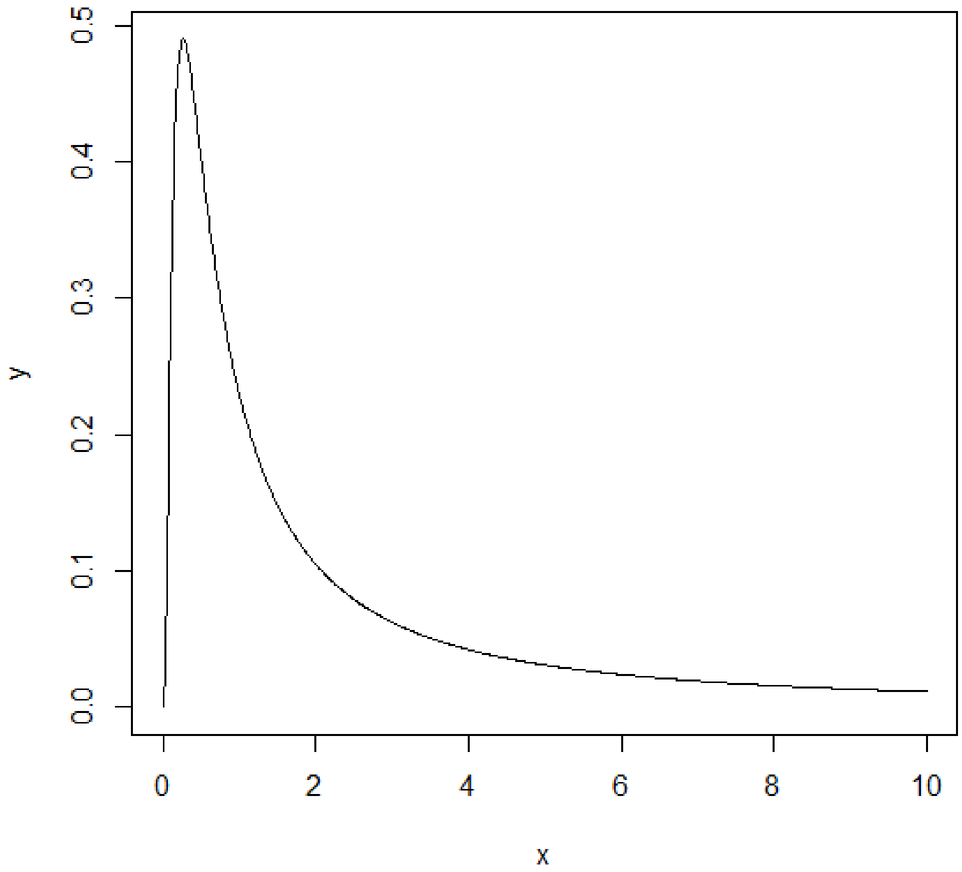
Two variances and the F distribution

$n_x = 5, n_y = 2$



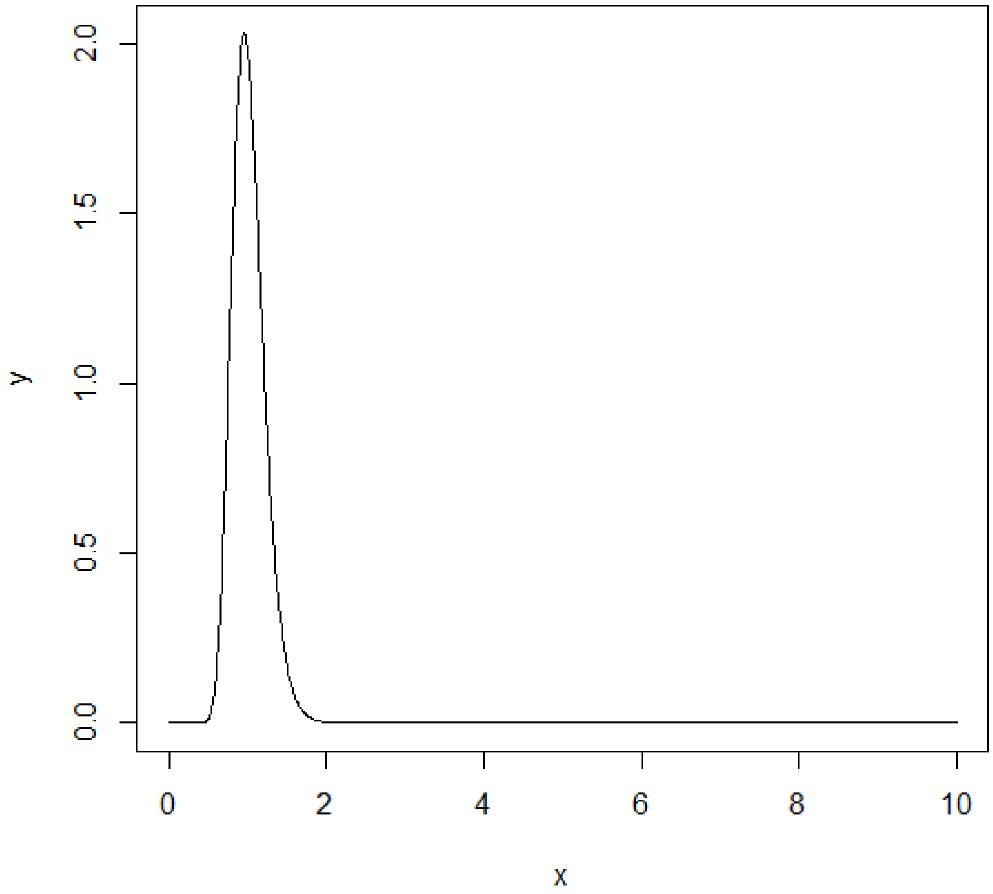
Two variances and the F distribution

$n_x = 10, n_y = 1$



Two variances and the F distribution

$n_x = 100, n_y = 100$



Two variances and the F distribution

- The F distribution can be defined as follows:
- $F_{n_x,n_y} = \frac{\left(\frac{X_x^2}{n_x}\right)}{\left(\frac{X_y^2}{n_y}\right)}$
- Where X_x^2 with n_x degrees of freedom and X_y^2 with n_y degrees of freedom are independent χ^2 random variables

Two variances and the F distribution

- Relating it back to $\bar{x}, s_x^2$, and the t-distribution

$$F_{n_x-1,n_y-1} = \frac{\left(\frac{\left(\frac{(n_x-1)s_x^2}{\sigma_x^2}\right)}{n_x-1}\right)}{\left(\frac{\left(\frac{(n_y-1)s_y^2}{\sigma_y^2}\right)}{n_y-1}\right)} = \frac{\frac{s_x^2}{\sigma_x^2}}{\frac{s_y^2}{\sigma_y^2}} = \frac{\frac{s_x^2}{s_y^2}}{\frac{\sigma_x^2}{\sigma_y^2}}$$

Two variances and the F distribution

- Thus allowing us to compare the variances of two populations

- We say $\frac{\left(\frac{s_x^2}{s_y^2}\right)}{\left(\frac{\sigma_x^2}{\sigma_y^2}\right)}$ follows F_{n_x-1,n_y-1}, the F distribution with $n_x - 1$ and $n_y - 1$ degrees of freedom

Summaries!

- Bunnies, Rabbits and the NY times
  - https://www.youtube.com/watch?v=jvoxEYmQHNM

Sampling Distribution for the Sample Proportion Summary

Shape of sample	Center of sample	Spread of sample
The shape of the distribution is bell shaped if $n * \rho \geq 15$ $\boldsymbol{and}$ $n * (1 - \rho) \geq 15$	$\mu_{\hat{p}} = \rho$	$\sigma_{\hat{p}} = \sqrt{\dfrac{\rho(1-\rho)}{n}}$

Sampling Distribution for the Sample Mean Summary

Shape, Center and Spread of Population	Shape of sample	Center of sample	Spread of sample
Population is normal with mean μ and standard deviation σ.	Regardless of the sample size n, the shape of the distribution of the sample mean is normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
Population is not normal with mean μ and standard deviation σ.	As the sample size n increases, the distribution of the sample mean becomes approximately normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

s_x^2 and the χ^2 (chi-squared) distribution

- The chi-squared density can be defined as follows:

$$f_{x(x)} = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} I(X > 0)$$

- Mean = n
- Variance = $2n$

χ^2 (chi-squared) distribution in R

- $P(X = x) = 0$ as the probability of any one value is always zero

- $P(X \leq x) =$ pchisq(x, n)

- $P(X \geq x) =$ 1 - pchisq(x, n)

- $P(x_1 < X < x_2) =$ pchisq(x_2, n) - pchisq(x_1, n)

s_x^2 and the χ^2 (chi-squared) distribution

- The chi-squared distribution can be defined as follows:

$$\mathrm{X}^2 = z_1^2 + z_2^2 + \cdots + z_n^2$$

- $\boldsymbol{X^2}$ **follows the chi-squared distribution with n degrees of freedom** where z_i are independent random variables that follow the standard normal distribution

The Chi-Squared Distribution

$$X_{n-1}^2 = \left(\frac{(n-1)s^2}{\sigma_x^2}\right) = \sum \left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2$$

- $\left(\frac{(n-1)s^2}{\sigma_x^2}\right)$ follows a χ^2 distribution with n-1 degrees of freedom

$\bar{x}, {s_x}^2$ and the Student's t-distribution

- The t density can be defined as follows:

$$f_{x(x)} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} I(X \epsilon \mathbb{R})$$

- Mean = 0
- Variance = $\frac{n}{n-2}$

$\bar{x}, {s_x}^2$ and the Student's t-distribution in R

- $P(X = x) = 0$ as the probability of any one value is always zero

- $P(X \leq x) =$ pt(x, n)

- $P(X \geq x) =$ 1 - pt(x, n)

- $P(x_1 < X < x_2) =$ pt(x_2, n) - pt(x_1, n)

$\bar{x}, {s_x}^2$ and the Student's t-distribution

- The t distribution can be defined as follows:

$$t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

- t_n **follows the t distribution with n degrees of freedom** where Z follows the standard normal distribution and χ^2 is chi-squared and divided by its degrees of freedom

$\bar{x}, {s_x}^2$ and the Student's t-distribution

$$t_{n-1} = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}} = \frac{\bar{x} - \mu_x}{(s_x/\sqrt{n})}$$

- $\frac{\bar{x}-\mu_x}{(s_x/\sqrt{n})}$ follows a t distribution with n-1 degrees of freedom
- Note the similarity to the z-score: the only difference here is that we estimate σ_x with s_x

Two variances and the F distribution

- The F density can be defined as follows:

$$f_{x(x)} = \frac{\left(\sqrt{\frac{(d_1 x)^2 d_2^2}{(d_1 x + d_2)^{(d_1+d_2)}}}\right)}{xB\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} I(x \geq 0)$$

- Mean = $\frac{d_2}{d_2-2}$ for d_2>2
- Variance =$\frac{2 * d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ for d_2 >4

Two variances and the F distribution in R

- $P(X = x) = 0$ as the probability of any one value is always zero

- $P(X \leq x) =$ pf(x, n_x, n_y)

- $P(X \geq x) =$ 1 - pf(x, n_x, n_y)

- $P(x_1 < X < x_2) =$ pf(x_2, n_x, n_y) - pf(x_1, n_x, n_y)

Two variances and the F distribution

- The F distribution can be defined as follows:
- $F_{n_x,n_y} = \dfrac{\left(\dfrac{X_x^2}{n_x}\right)}{\left(\dfrac{X_y^2}{n_y}\right)}$
- Where X_x^2 with n_x degrees of freedom and X_y^2 with n_y degrees of freedom are independent χ^2 random variables

Two variances and the F distribution

$$F_{n_x-1,n_y-1} = \frac{\left(\frac{s_x^2}{s_y^2}\right)}{\left(\frac{\sigma_x^2}{\sigma_y^2}\right)}$$

$\frac{\left(\frac{s_x^2}{s_y^2}\right)}{\left(\frac{\sigma_x^2}{\sigma_y^2}\right)}$ follows, the F distribution with $n_x - 1$ and $n_y - 1$ degrees of freedom