ECON 253: INTRODUCTION TO ECONOMETRICS

Sebastian Anti, Assistant Professor of Economics
Bryn Mawr College

Admin

- Stata troubles?
- Questions from last time

Lecture 2: Statistical Foundations

Objectives

- **Introduction to basic statistical concepts:**
  - Probability
  - Random variables
  - Expected values
  - Quantifying uncertainty over estimates (interval estimation)
  - Testing hypotheses
  - Variances
  - Conditional distributions

Motivational Exercise

- **Consider this challenge:**
  - You are a beer brewer and concerned about quality control. You are worried a certain amount of your beer is contaminated and you want to know what percent of your bottles are undrinkable. Testing beer requires that you destroy it afterwards. **How do you measure how much of your beer is contaminated?**
- **Consider this challenge:**
  - You are the US Secretary of Education. You want to know what percent of high school graduates in the US read at a college level. Testing all recent high school graduates is unfeasible. **How do you measure this?**

Motivational Exercise (cont.)

- **How about this challenge:**
  - You are a planetary scientist measuring how much of the Earth's surface is water and how much is land (imagine you are doing this before anyone else has done it). Inventorying each point on the globe is prohibitively time consuming. How do you do this?
- **Turns out each of these is a version of the same problem:** How does one measure something when it is prohibitively costly or logistically impossible to actually measure it?

Earth Surface Measurement Problem

- **Let's take this last one and use Statistics to answer it**
  - (Yes we're actually using real statistics here!)
- **The way this is going to work:**
  - We'll throw the globe (with a decent amount of spin) from student to student
  - When you catch it (or pick it up), tell me whether the tip your index finger on your right hand is over water or land
  - I will record the data this generates on the board

What is Statistics?

- **Basic definition:**
  - The practice of collecting and analyzing numerical data

- **General idea behind statistics:**
  - You want to know something about a **population** like the population's average income or education level, called a "**parameter**"
    - Note this is different use of the term "parameter" than in the econometrics context
  - But you can't measure this for every individual in the population (too expensive perhaps)
  - So you take a **sample** (a subset of individuals) from that population that you think is representative of the larger population
  - Find the measurement for the sample, called a "**statistic,**" and then use that as an estimate for the population's parameter
    - But how confident are we that this statistic is a good measure of the population? Statistics provides useful ways to figure this out.

Why Do We Need Statistics?

- **Statistics is the foundation of Econometrics**
  - Econometrics uses data from samples of larger populations
- **Using a sample to estimate relationships in a larger population requires us to contend with the uncertainty this introduces**
  - Never sure if the sample is a true reflection of the population
  - Possible your statistics are just the product of you selecting a bad sample rather than a measure of your population
- **Statistics provides extremely useful tools for dealing with that uncertainty**

Preliminary Definitions

- **Underlying probability theory and statistics is the idea of a "random chance experiment." Examples:**
  - Flipping a coin
  - Throwing die
  - **Example in Econ:** Choosing a person at random to include in a sample and measure some economic characteristic like her/his/their wage
- **A "sample space" (S) of a random experiment like this is the set of all possible outcomes:**
  - In a coin toss: $S = \{H, T\}$
  - In a die throw: $S = \{1, 2, 3, 4, 5, 6\}$
  - In a random sample it's all possible wage values in a population

Preliminary Definitions

- **An "event" is a subset of the sample space:**
  - In a coin toss: $\{H\}\ \{T\}$
  - In a die throw: $\{Even\ Number\}\ \{> 2\}$
  - In picking a worker: $\{wage > 10\ per\ hour\}$
- **Two events are "mutually exclusive" if they have no outcomes in common:**
  - In coin toss: $\{H\}\ \{T\}$
  - In a die throw: $\{1\}\ \{> 2\}$
  - In picking a worker: $\{wage < 10\ per\ hour\}\ \{wage > 20\ per\ hour\}$

Preliminary Definitions

- **A set of events is "exhaustive" if it contains all elements of a sample space:**
  - In a coin toss: $\{H\}\ \{T\}$
  - In a die throw: $\{1,2,3\}\ \{4,5,6\}$
  - In picking a worker: $\{wage > 10\ per\ hour\}\ \{wage \leq 10\ per\ hour\}$

Definition of Probability

- **The probability of an event A, denoted $P(A)$, is the proportion of times event A occurs in repeated trials of the random chance experiment**
  - $P(\mathit{Coin\,Toss} = H) = ½$
  - $P(\mathit{Draw\ Person\ with\ wage} > 100\ \mathit{per\ hour}) =$ Share of population earning greater than 100 per hour
- **Concept of repeated trials is important**
  - This means we are doing the experiment many, many (an infinite) number of times
  - In limited repetitions there's a good chance we don't get heads exactly half the time when tossing a coin

Properties of Probabilities

- **$P(A)$ is a real-valued function that has the following properties:**
  - It takes values between 0 and 1, inclusive ($P(A) \in [0,1]$)
  - If $A_1, A_2, \ldots A_N$ is an *exhaustive* set of events, then $P(A_1 \text{ } or \text{ } A_2 \text{ } \ldots or \text{ } A_N) = 1$
  - If $A_1, A_2, \ldots A_N$ is a *mutually exclusive* set, then $P(A_1 \text{ } or \text{ } A_2 \text{ } \ldots or \text{ } A_N) = P(A_1) + P(A_2) + \cdots + P(A_N)$

Properties of Probabilities

- **These properties can help us answer the following questions**
  - $P(Wage < 10\ USD\ or\ Wage \geq 10\ USD) =?$
    - Answer: 1
  - If $P(Die\ shows\ 3) = 0.2$ and $P(Die\ shows\ 4) = 0.2$ then what is $P(Die\ shows\ 3\ or\ 4)$?
    - Answer: 0.4
  - If $P(Wage > 15\ USD) = 0.4$ then what is $P(Wage \leq 15\ USD)$?
    - Answer: 0.6

Discrete and Continuous Random Variables

- **A "random variable" is a function or process that assigns a numerical value to the outcomes of a random or chance experiment or activity**
  - In coin toss: a random variable $Y = 1$ if heads and $Y = 0$ if tails
  - In picking a worker: Define $Y = \ person's \, wage$
- **A "discrete random variable" takes an integer value**
  - Example: years of education
- **A continuous random variable takes on any real value**
  - Usually this falls in some range

Returning to our Globe Example

- What is the "parameter" we are after?
- What is the "statistic" we measured?
- What was the "random chance experiment" we did? What was "random" about it?
- What was the "sample space" of that experiment?
- What are the "events" in this sample space? Are they mutually exclusive?
- How would we have been misled if we had not done "repeated" trials after the first 3?

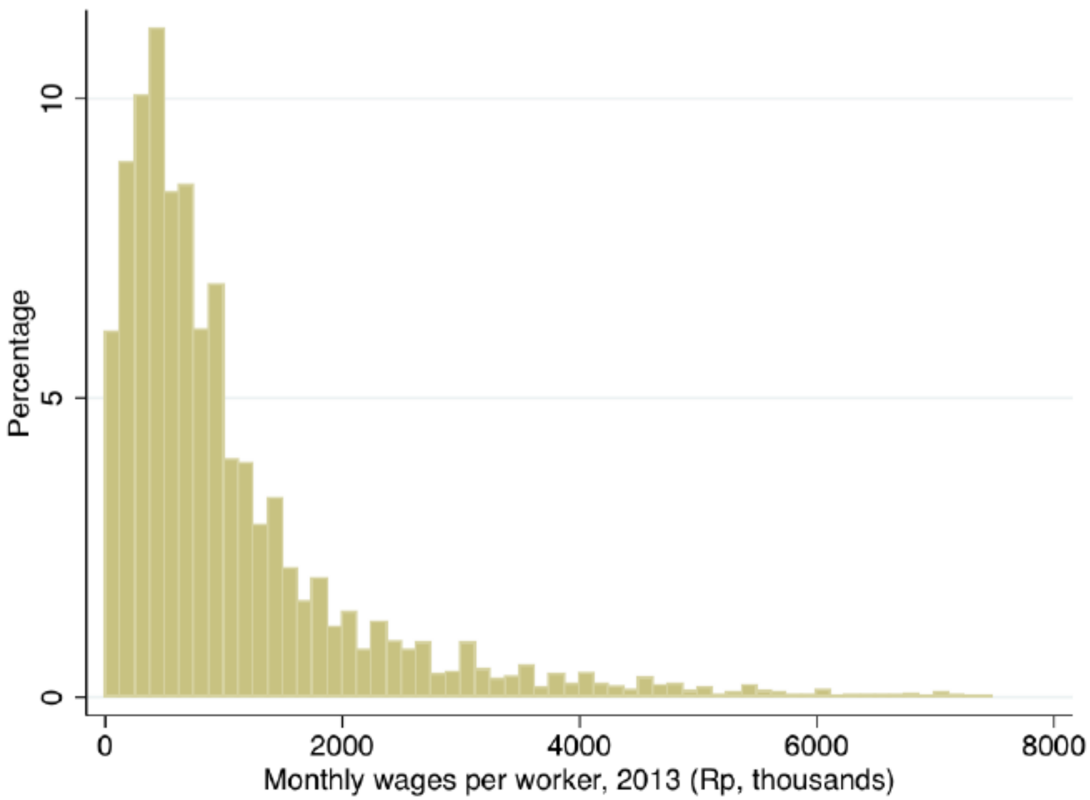
Returning to our Globe Example

- Based on our data, what is $P(A_{water})$?
- What is the value of $P(A_{land}) + P(A_{water})$? Why?
- What is the "random variable" in this example?
- Is this random variable discrete or continuous?

Distributions of Random Variables

- **Random variables vary across individual observations**
- **The "distribution" of a random variable tells us information about that variation**
  - All random variables have distributions
- **It is usually shown visually by a "density histogram"**
  - This is just a bar chart where the x-axis contains intervals within the sample space and the y-axis is the the percentage of observations that take values within each interval
  - The total area covered by the histogram equals 1

Density Histogram

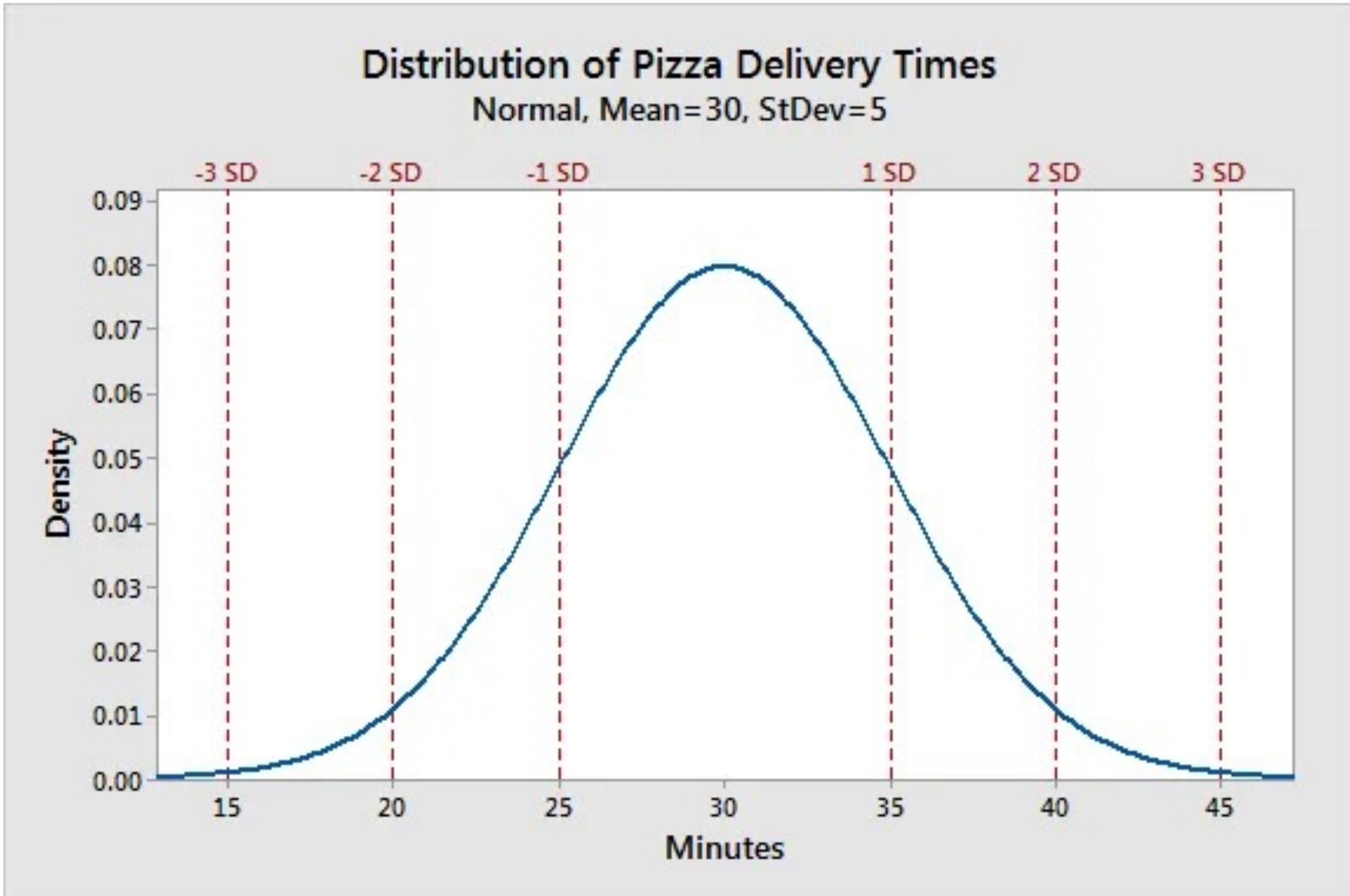


- **Here's an example from India**
  - Note the rich amount of information this is telling us about the probabilities one would observe certain wages
- **What is the density histogram of our globe example random variable?**

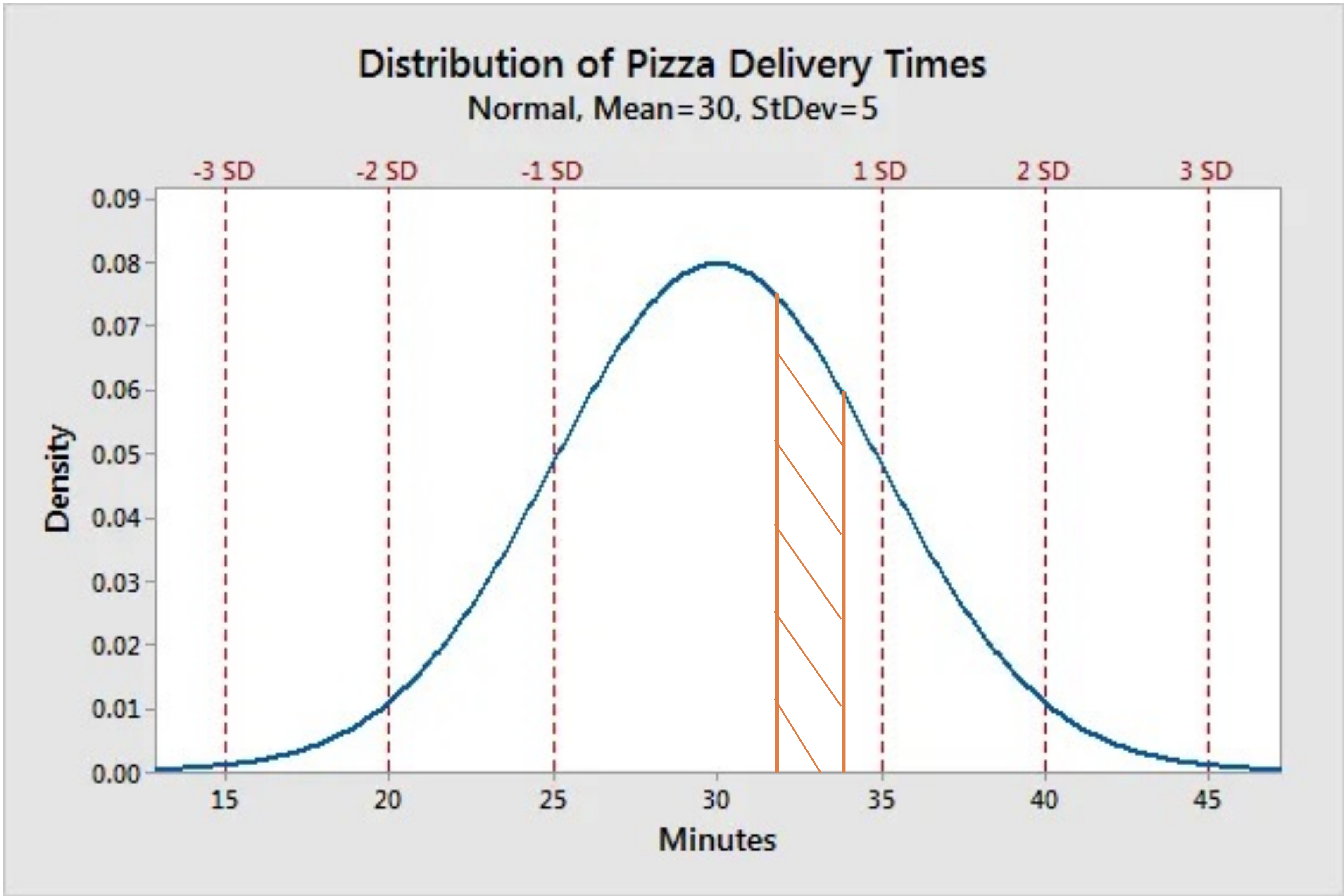
Distributions of Random Variables

- **Understanding the distribution of a random variable allows one to understand many important qualities of it**
- **A common shape that many random variables take on is the "normal distribution" (but far from the only one)**
  - Also known as the "bell curve" distribution
- **Characterized by large mass of observations in the center intervals of the sample space, with symmetric "tails" at the lower and higher values of the sample space**

Example of a Distribution



Example of a Distribution



Quantifying the Characteristics of a Distribution

- **There are four (really two for our purposes) qualities to focus on in describing a distribution of a random variable:**
  1. **Its <u>central tendency</u>:** identification of the central value of the distribution
  2. **Its <u>dispersion</u>:** the extent to which the distribution is stretched or squeezed
  3. **Its <u>kurtosis</u>:** qualities of the tails (not relevant for our purposes – just mentioned for sake of completeness)
  4. **Its <u>skewness</u>:** symmetry (not relevant for our purposes – just mentioned for sake of completeness)

Measures of Central Tendency

- **There are three traditional measures of a random variable's central tendency:**
  1. **The mean (AKA the "average"):** $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$
  2. **The median:** the value separating the lower half from the upper half of the distribution (if N is even, the median is the mean of the two middle values)
  3. **The mode:** the most common value in the distribution

- **Note: the mean of a normal distribution is found at its peak**
- **What are these values in our globe random variable?**

Measures of Dispersion

- **There are three measures of dispersion**
  - **The range:** the difference between the largest and smallest values of the distribution
  - **The variance:** the mean of the squared deviations from the mean: $\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = Var(x)$
  - **The standard deviation:** the square root of the variance: $\sigma_x = \sqrt{Var(x)} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}}$
- **Measures of dispersion and central tendency are often called "descriptive statistics"**

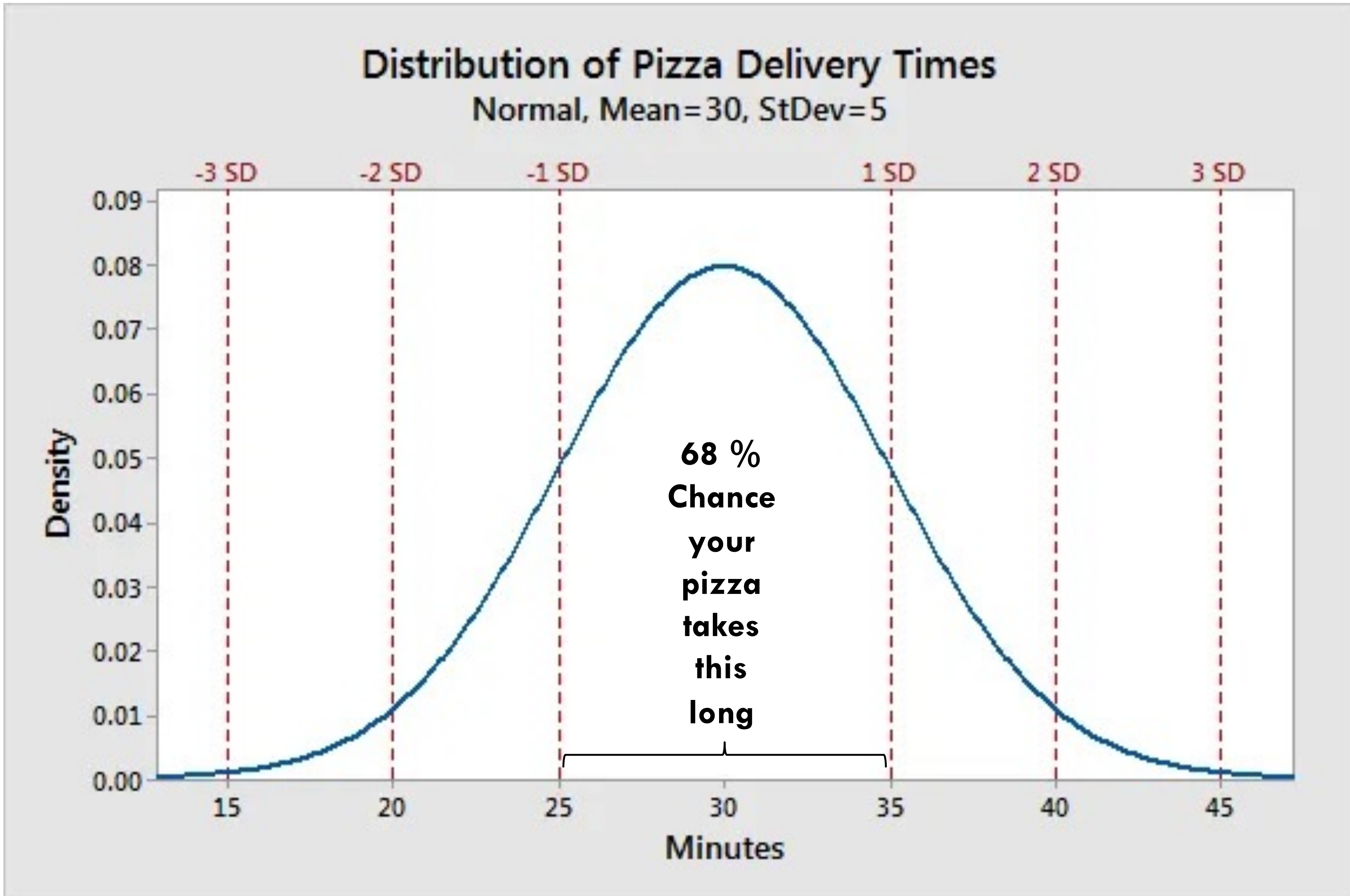
Measures of Dispersion

- **Standard Deviation benefits and interpretation:**
  - Often a nice way to express information about variance in the same units as the original variable
  - Tells us the average extent to which the variable's values lie above or below its mean
- **If the variance is zero, the "random variable" is not really random, it is just a "constant," "scalar," or "nonrandom variable"**

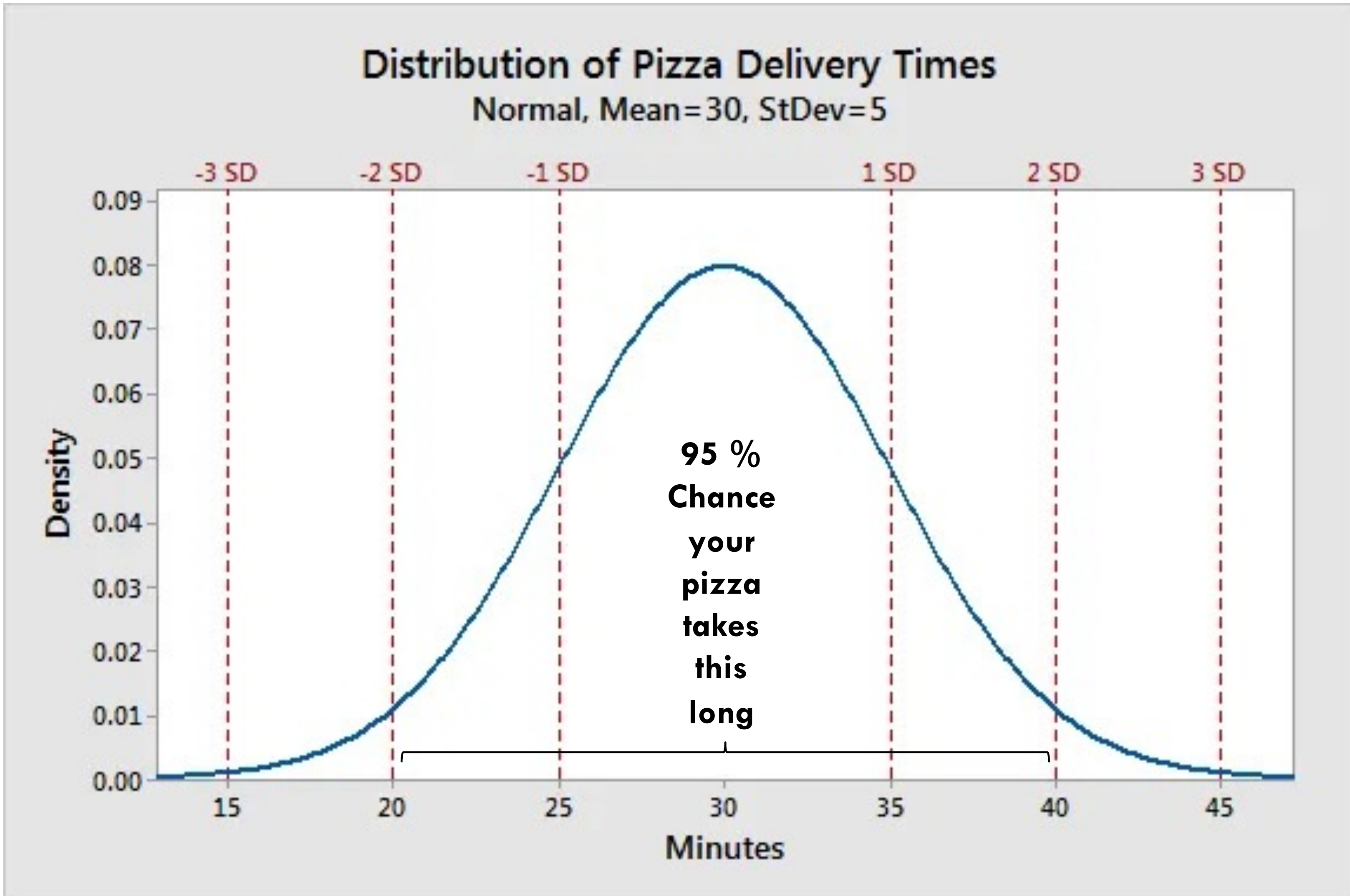
Nice Properties of the Normal Distribution

- As mentioned before, symmetric around the mean
- About 68% of its mass is within 1 standard deviation of its mean
  - This means about 68% of observations of any random variable that has a normal PDF take values within 1 standard deviation of its mean
- About 95% of its mass is within 2 standard deviations of its mean
- About 99.7% of its mass is within 3 standard deviations of its mean

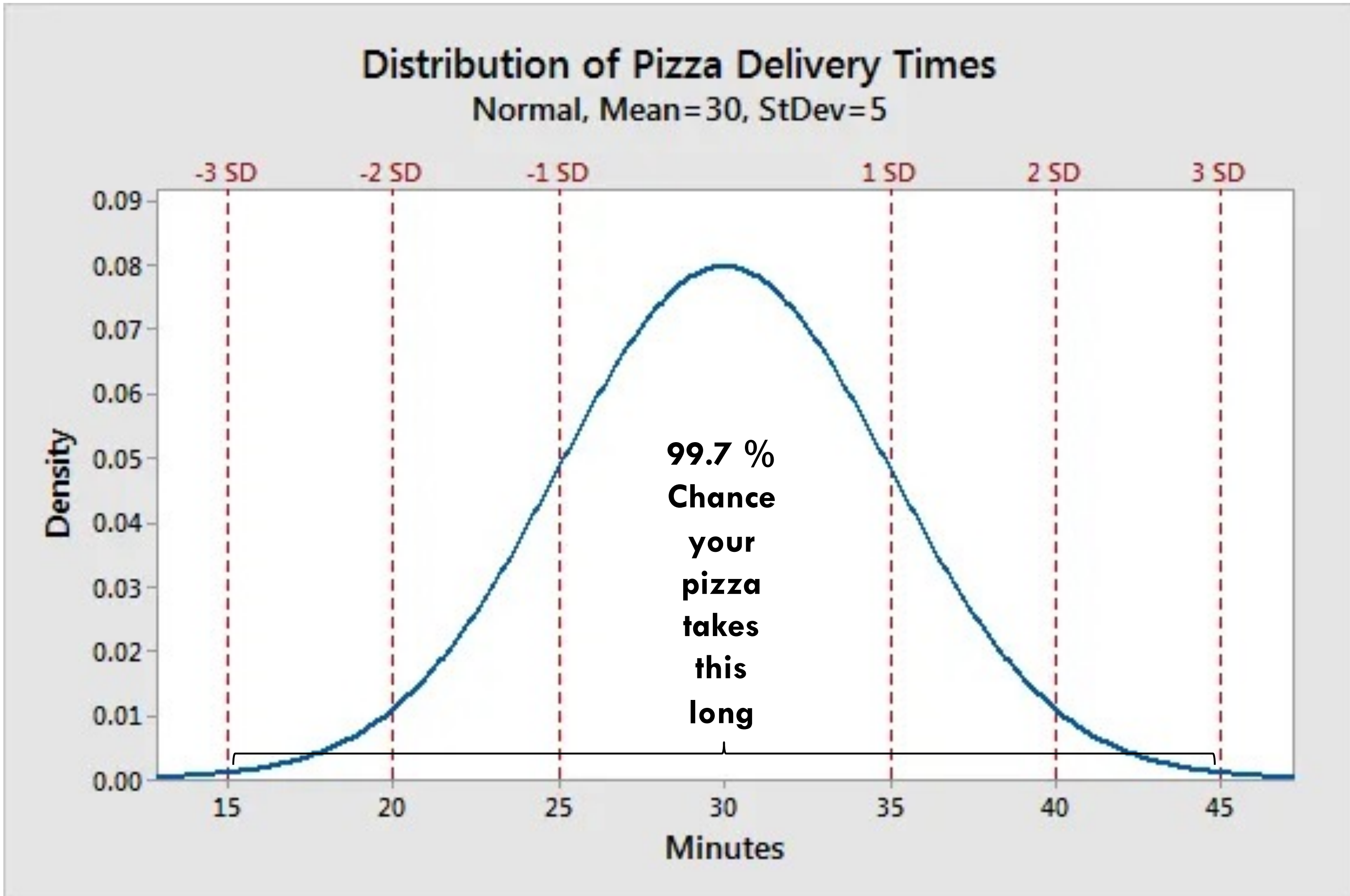
Example of a Distribution



Example of a Distribution



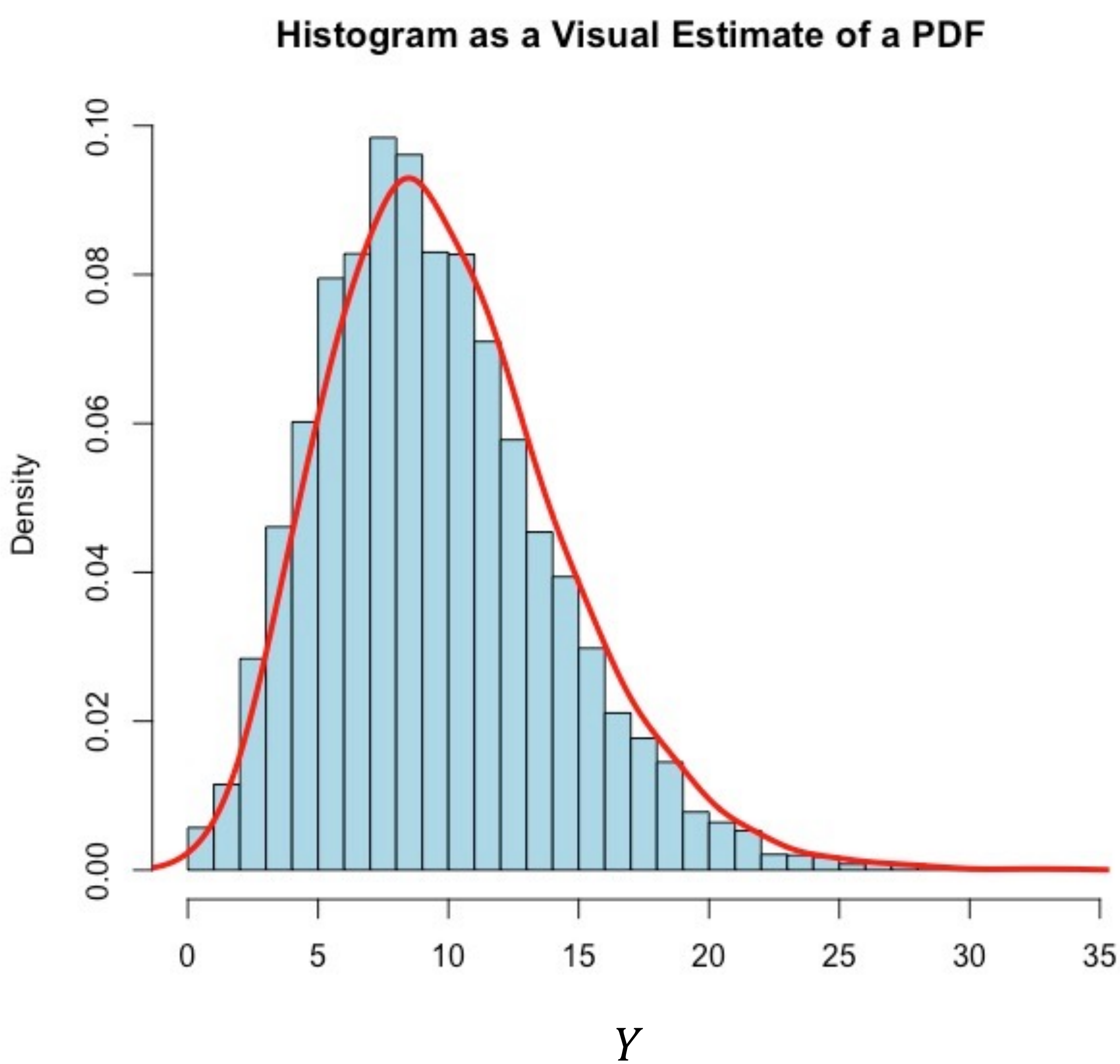
Example of a Distribution



Probability Density Functions

- **The distribution of a random variable Y is summarized by a function called the "Probability Density Function"**
  - a statement of the possible values that Y might ever take on, along with a statement of the probability of Y taking on each of those values in that range
- **We don't observe these directly,**
  - but we can often learn about them by studying random samples
    - Density histograms provide a visual estimate of the PDF of a random variable

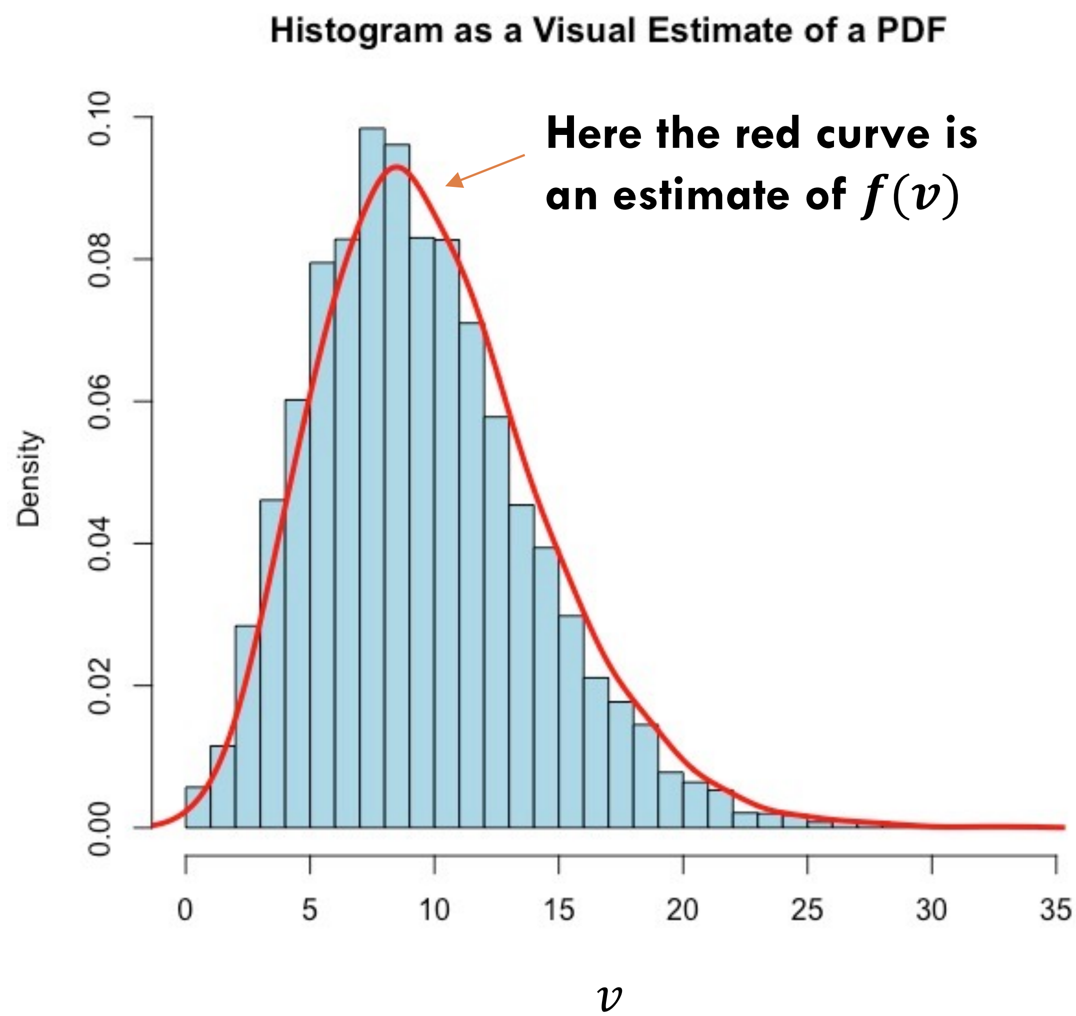
Histogram as an Estimate of a PDF



- **What does this PDF suggest is the "sample space" of the this random chance experiment?**
  - Recall: A "sample space" (S) of a random experiment is the set of all possible outcomes

Probability Density Functions

- Easier to make sense of PDF for a continuous random variable by graphing it
- Which event has a higher likelihood of occurring, an event between 0 and 5 or an event between 10 and 15?
  - 10 and 15 b/c more events occur in this range



Example of a Continuous Probability Density Function

- **A normal random variable with a mean, μ, and a variance of σ^2 has the PDF,**

Mean of v

$$f(v) = \frac{1}{\sqrt{2\pi\sigma}} \times exp\left(-\frac{(v-\mu)^2}{2\sigma^2}\right) \qquad (2)$$

Variance of v

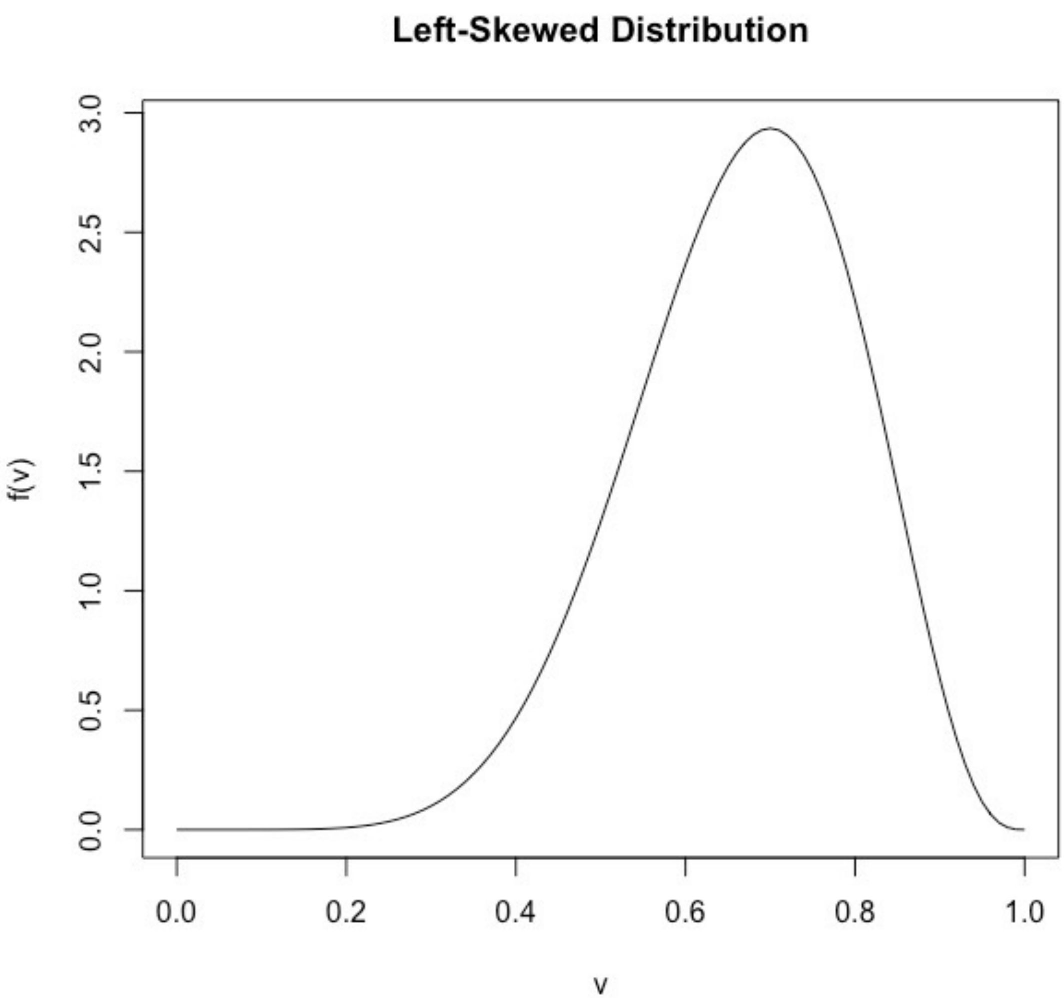
Principles of PDFs and Continuous Random Variables

- **Fact: The area under the curve of the PDF must equal 1:**
  - Why?: Because the area under the curve contains the exhaustive sample space
  - Implies that if $P(y \leq a) = 0.75$, then $P(y > a) = 0.25$

Question from Last Time

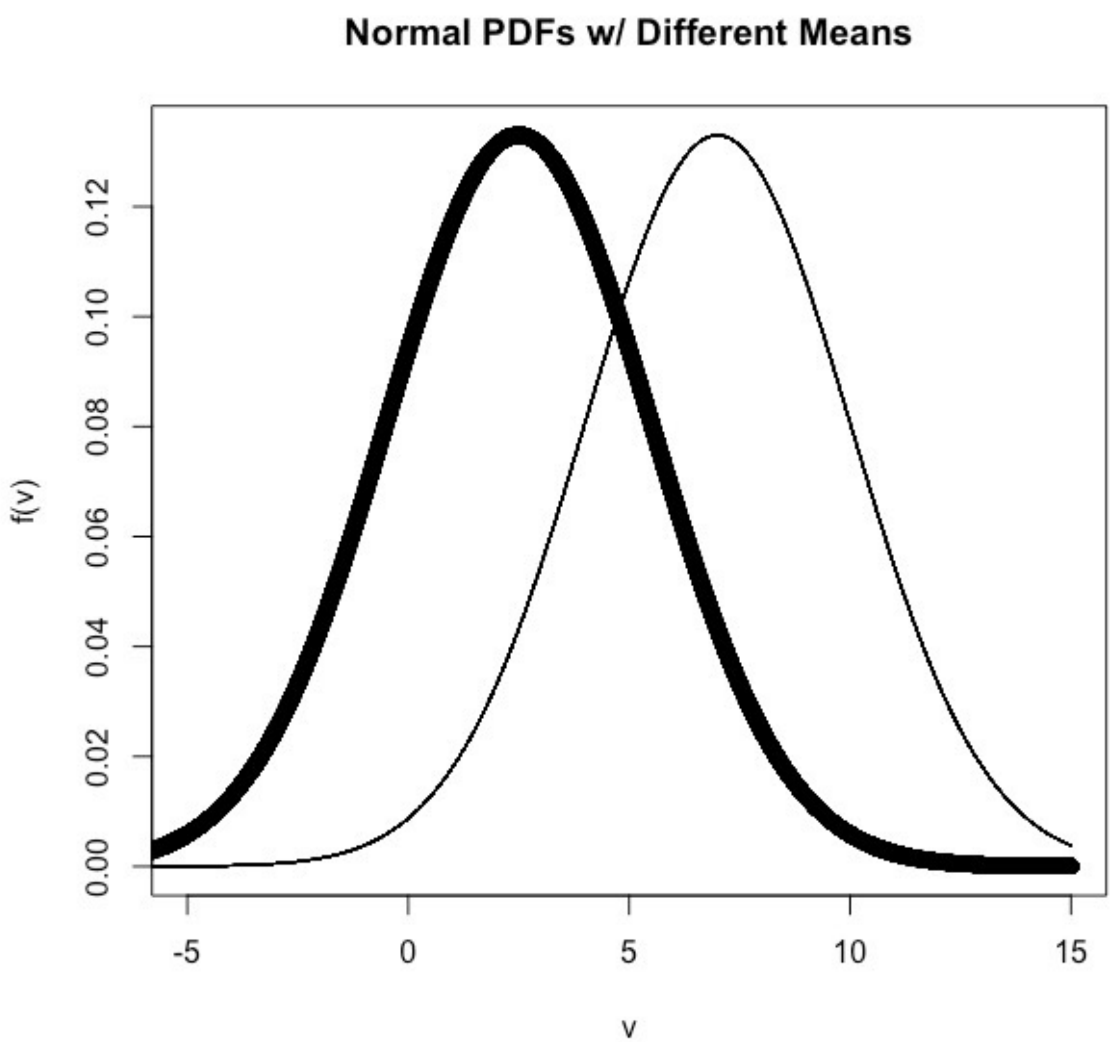
- **Why use a square of the deviation from the mean in calculating a sample's variance? Why not use absolute value instead?**
  - Recall: $\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = Var(x)$
- **Best Answer:** It is easier to work with analytically (easier to differentiate and integrate) than absolute values. So statisticians find it more convenient to use in proving theorems etc.

Measures of Central Tendency (Again)



- The mean, or expected value, of a random variable with a skewed distribution is to the left (right) of the peak if skewed left (right)

Measures of Central Tendency (Again)



- These two distributions have different means

Measures of Central Tendency (Again)

- **What can we learn from an expected value?**
  - Tells us something about the location of the main non-zero parts of the PDF along the v axis
  - In common parlance in econometrics it tells us where the "mass" is of the distribution
  - A general sense of the most common values the random variable takes

Importance of PDFs

- **When we are studying a random variable of an important measure in the real world, the PDF provides "<u>real information</u>" about an underlying reality about the world**
- **In other words, it provides its "distribution":**
  - Central tendency of the random variable
  - Dispersion around the central tendency
  - "Skewness" (measures of the symmetry of the PDF)
  - "Kurtosis" (measures of the tails of the PDF)

Statistical Inference

- **Now, let's say you take a sample and measure a mean value of interest to you. Good for you!**
- **How confident are you that this statistic is a "good" estimate of the population parameter of interest?**
  - Remember, you can never actually know what the population parameter's value is
- **Statistics provides us a structured way of quantifying how sure we are that our estimate is close to the population parameter**
  - Making claims about this is called "statistical inference"

Confidence Intervals

- **The standard approach for this is to construct a "confidence interval"**
  - This provides a range of numbers in which one expects to see the population parameter to be found, given the qualities (central tendency and dispersion) of the distribution
- **Intuition based on:**
  - **Remember, that it is a fact of the normal distribution that 95 percent of observations fall within two standard deviations of the mean**
  - **We'll also need to invoke the Central Limit Theorem (CLT)**
    - This says that the distribution of means from repeated samples from the population (with replacement) is normal

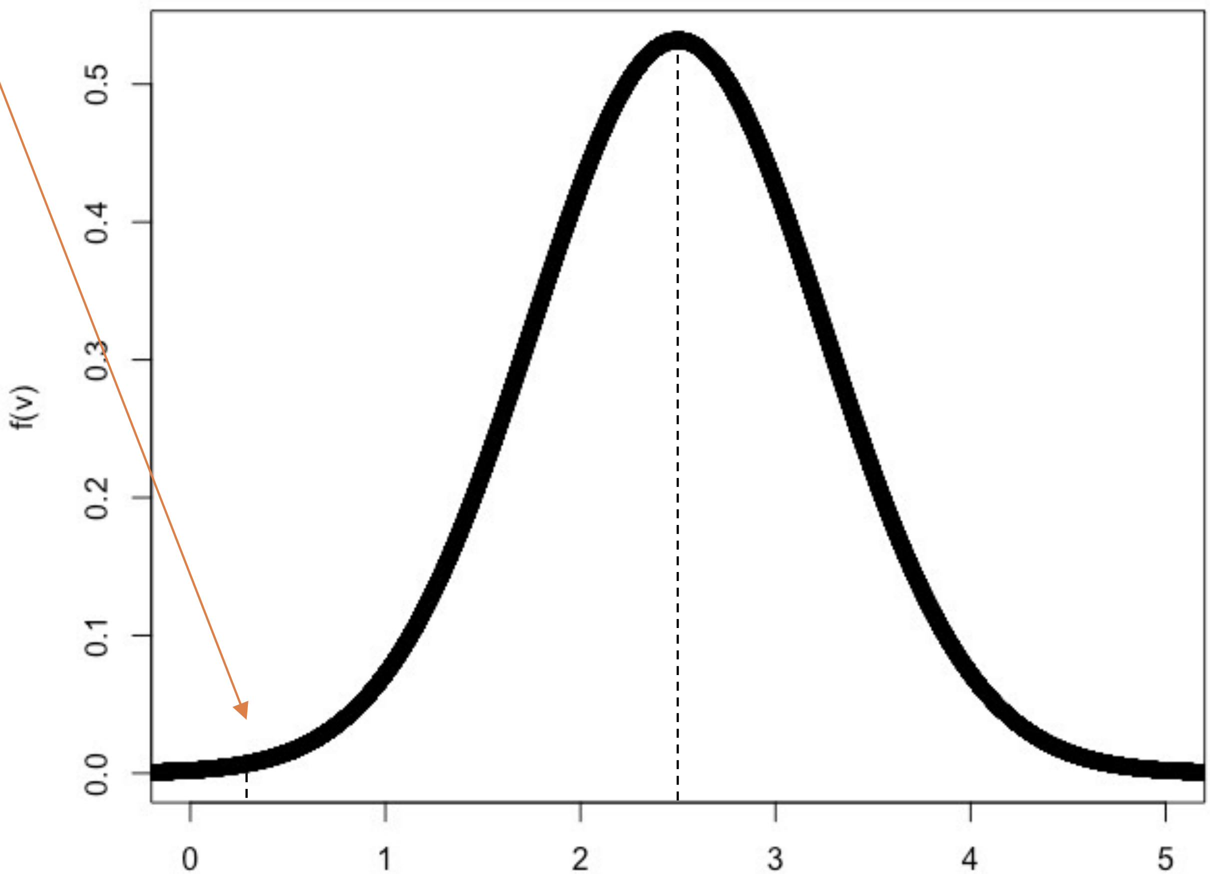
The Central Limit Theorem (CLT)

- **What the CLT says is:**
  - If you repeatedly sampled from the population and calculated $\bar{x}$ over, and over again, then…
  - the distribution you'd get for $\bar{x}$ would be normal, with a the mean equal to the population mean μ_x and its standard deviation would be equal to $\frac{\sigma_{\bar{x}}}{\sqrt{N}}$, where $\sigma_{\bar{x}}$ is the SD of the sampling distribution
- **In notation:** $\boldsymbol{\bar{x} \sim N(\mu_x, \frac{\sigma_{\bar{x}}}{\sqrt{N}})}$
- **<u>This is an extremely powerful result</u>**

What the CLT Means

Remember though, in practice we only see $\overline{x}$ once. How can we be sure we aren't observing this situation here?

Distribution of Repeated Samples calculating $\overline{x}$



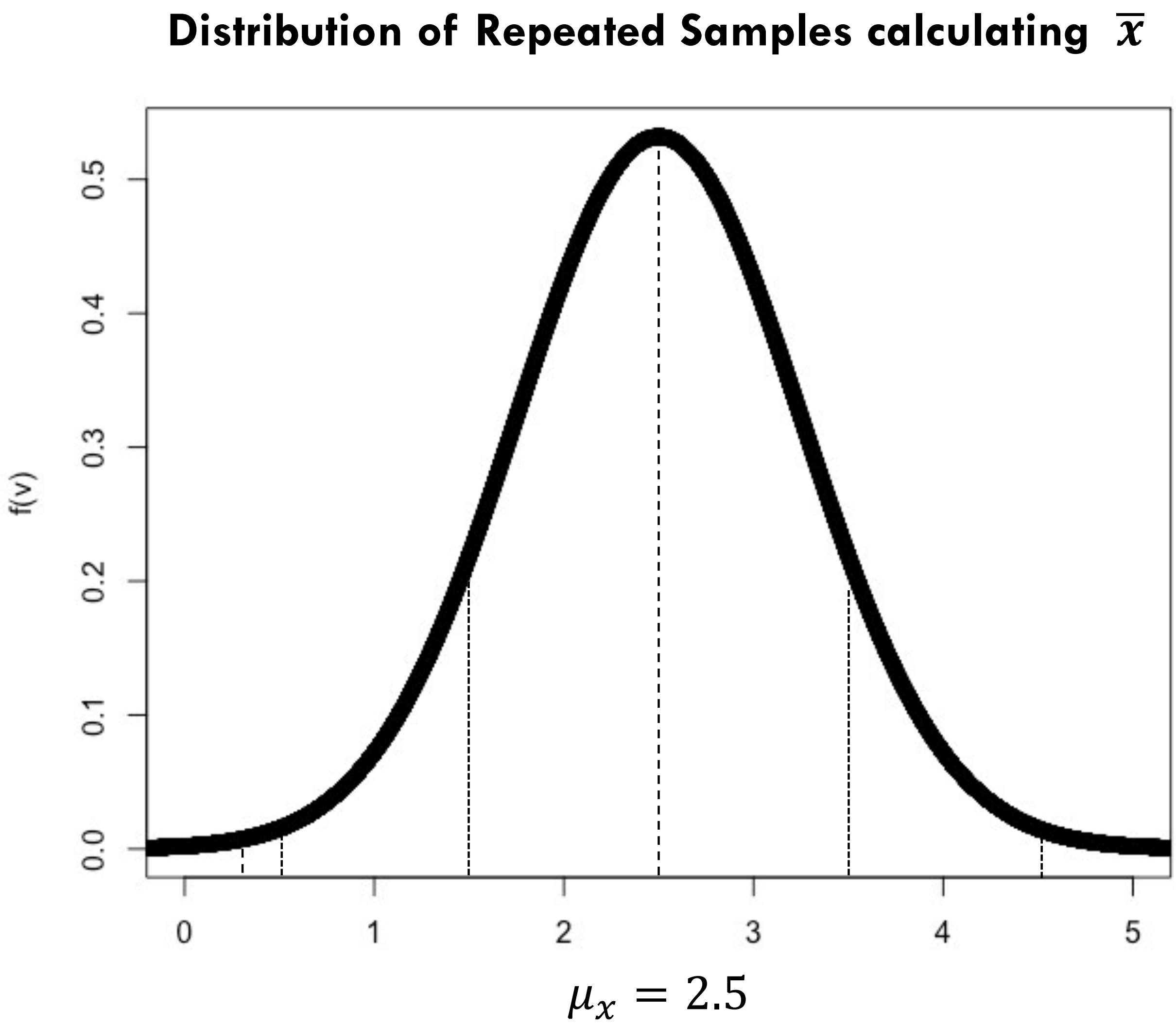
$\mu_x = 2.5$

Calculating the Z-score

- **We'll use the known properties of the normal distribution to our advantage here**
  - Remember, we know that 68% of observations fall within 1 SD of the mean, 95% fall within 2, and 99.7 within 3

What the CLT Means

- **If we know μ_x** and the SD, we can look at our observed value and say, "well it's outside of 2 SDs of the mean, therefore there is at most a 2.5% chance we'd see a value like this."



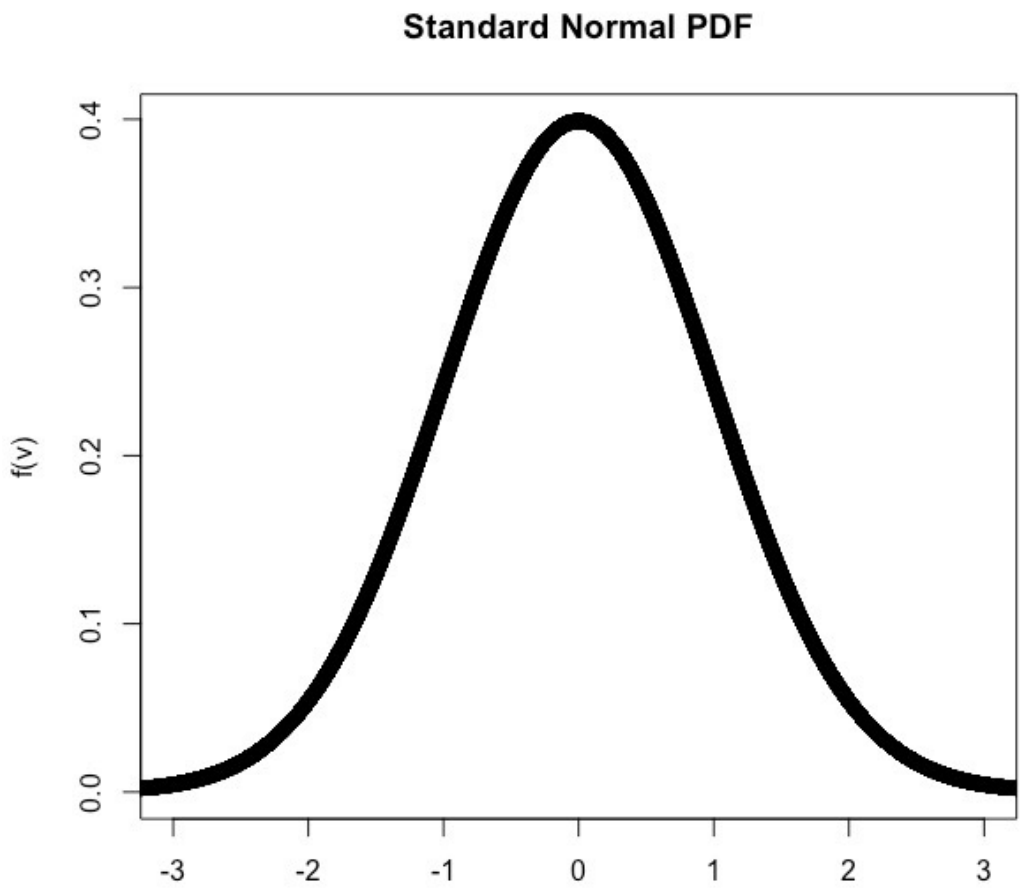
Calculating the Z-score

- **We'll use the known properties of the normal distribution to our advantage here**
  - Remember, we know that 68% of observations fall within 1 SD of the mean, 95% fall within 2, and 99.7 within 3
- ***In fact*, we can say even more than this because we know the probabilities associated with the normal distribution for values at other increments away from the mean**
  - Not just 1, 2, and 3 SDs away, but all values
  - Summarized in the z-table in the back of Wooldridge

Calculating the Z-score

- **To make a precise statement, we calculate the "Z-score":** $z = \frac{\bar{x} - \mu_x}{\sigma_x}$
  - This provides us the number of standard deviations an observation is away from the mean of the distribution
    - This "standardizes" the observation
- **The z score has it's own distribution called the "standard normal" distribution**
  - **The Standard Normal Distribution is important because:**
    - Any probability statement about a normal random variable can be transformed into a probability statement about a standard normal random variable
    - So with this transformation, we can understand the PDF of any normally-distributed random variable by looking at the Standard Normal PDF

The Standard Normal PDF



Z
(# of SDs above or below the mean)

- **A type of normal distribution with,**
  - $\mu = 0$ and
  - $\sigma^2 = 1$ (SD=1)
- **The probability that an observation from a normal random distribution is smaller than the observation is provided in a table**
  - Wooldridge Table G.1, p. 784

Standard Normal Probabilities

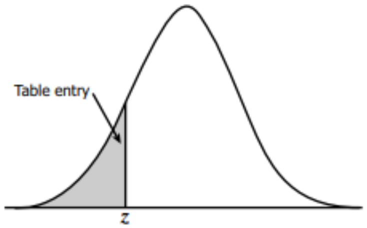


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

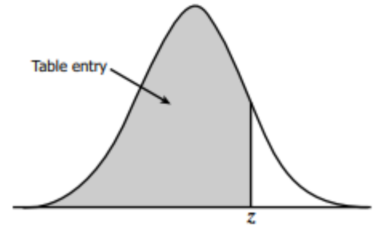


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

What to do with the Z Score

- **Once you calculate the z score, you then look up in a Z Table (Table of Wooldridge Table G.1, p. 784) what the probability is that the observed value you have is away from the mean**
  - Example from above: $z = \frac{0.25-2.5}{1} = -2.25$

Standard Normal Probabilities

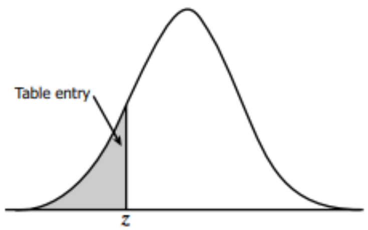


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
$\boxed{-2.2}$	.0139	.0136	.0132	.0129	.0125	$\boxed{.0122}$	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

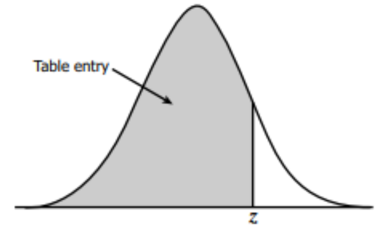


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Probability you would observe a value as large as 0.25 = 0.0122 or 1.22%

More Likely, You don't Know μ_x and σ_x

- **We're trying to use the sample to figure these things out after all!**
- **Almost certainly you don't know them, and calculating the Z-score isn't possible**
- **Don't despair. There is a solution to this**
  - By the CLT, we can use $\frac{\sigma_{\overline{x}}}{\sqrt{N}}$ as an estimate for σ_x
  - But we still don't know μ_x, so we'll use our knowledge of the normal distribution to come up with a range of values in which we are confident μ_x exists given our observed $\bar{x}$
    - This is another way of saying a "confidence interval"

Confidence Intervals

- **With this, we can construct a confidence interval:** a range around our sample's point estimate within which we are reasonably confident the population mean is
- **Remember that with a normal distribution, 95% point estimates will be within 2 SDs of the population mean**
- **So what we can say is that, we are 95% confident that the mean we've calculated, $\bar{x}$, is within 2 SDs of the population mean**

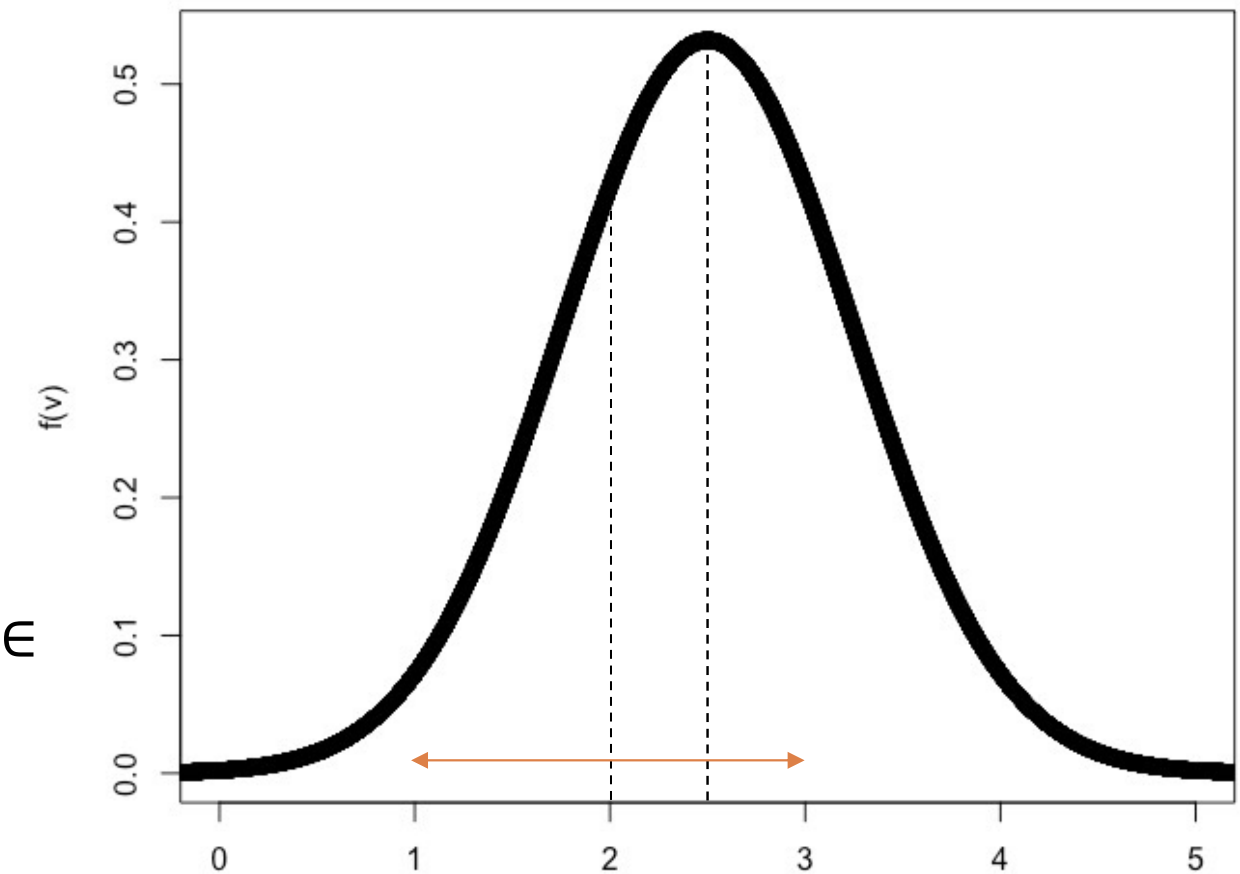
Confidence Intervals

- **Another way of saying this is that the population mean is within 2 SDs of $\overline{x}$, 95% of the time**
- **So we can say that we are 95% confident that $\boldsymbol{\mu_x(95\%\,Conf.) \in [\overline{x} - 2\times \frac{\sigma_{\overline{x}}}{\sqrt{N}}, \overline{x} + 2\times \frac{\sigma_{\overline{x}}}{\sqrt{N}}]}$**
  - $2\times \frac{\sigma_{\overline{x}}}{\sqrt{N}}$ is known as the "margin of error". Here it is 95% Confidence (5% margin of error)
    - If we wanted to be 99.7% confident how would this change?
- **This is very powerful!**
  - Even though we don't know μ_x, we are able to have a very strong statement about what it is likely to be!

Example

- **So perhaps $\overline{x} = 2$. Let's say $N = 100$, and $\sigma_{\overline{x}} = 5$**
- **What is the 95% confidence interval for the population mean?**
  - $\mu_x(95\%\ Conf.) \in [\bar{x} - 2\times\frac{\sigma_{\bar{x}}}{\sqrt{N}}, \bar{x} + 2\times\frac{\sigma_{\bar{x}}}{\sqrt{N}}]$

Distribution of Repeated Samples calculating $\overline{x}$



$\mu_x = 2.5$

Hypothesis Testing

- A question that might easily come up that statistics can help us to answer is: "given our observed $\overline{x}$, how confident are we that μ_x is some (interesting) value?"
- Hypothesis testing let's us answer this type of question
- Hypothesis testing is highly related to confidence intervals

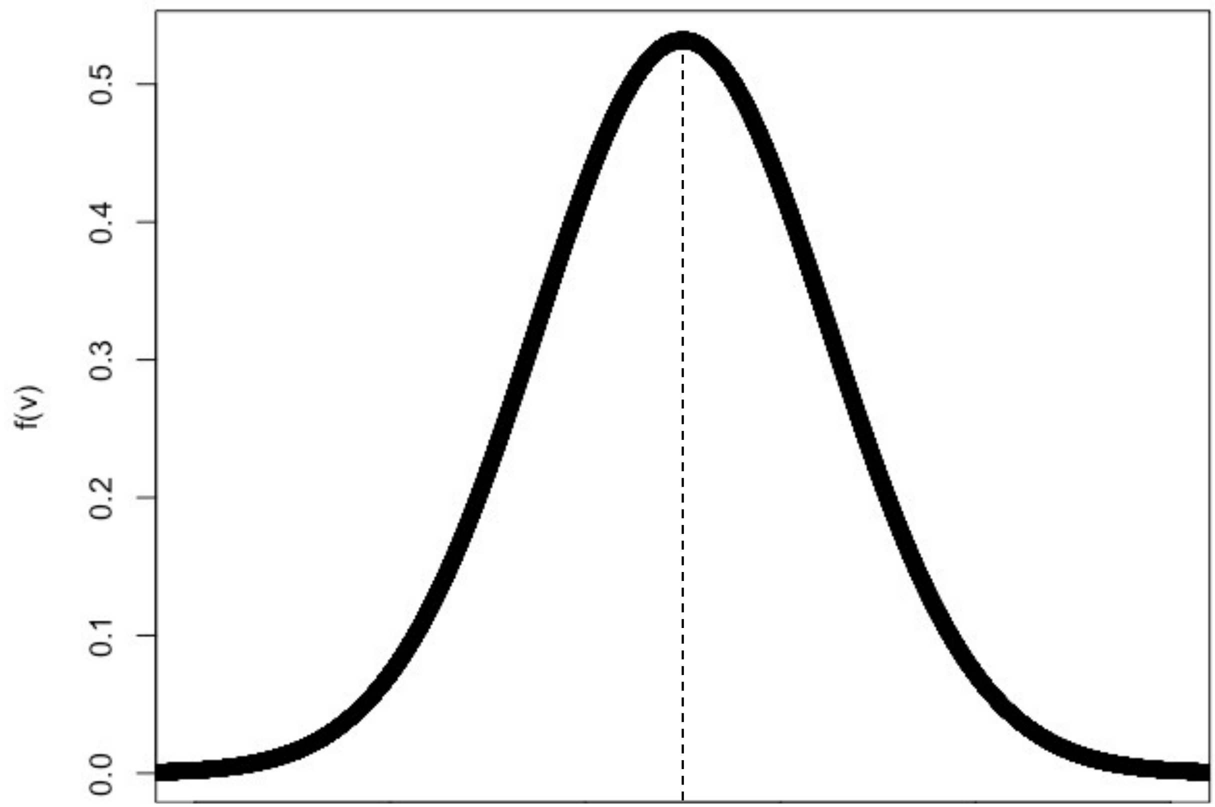
Hypothesis Testing

- **To accomplish this, you need to specify the null (H_0) and alternative (H_A) hypotheses**
  - The null hypothesis is a statement about the true value of the parameter of interest
  - The alternative hypothesis is a statement covering all the ways in which the null hypothesis could be false
- **We want to find out whether the data provide sufficient evidence to reject the null hypothesis with any level of confidence (usually 95% confidence)**
  - A way to answer this is to suppose the null is true. What would that mean? It would mean the following is true (next slide)
- **Example: Given the data collected on Pennsylvania individuals, how confident are we that the population's mean income is 70,000 USD?**
  - H_0: The mean income in Pennsylvania, $\bar{y}_{PA}$, is 70,000 USD annually or $\bar{y}_{PA} = 70{,}000$
  - H_A: $\bar{y}_{PA} \neq 70{,}000$

Hypothesis Testing Example

- **Let's say the sample mean equals $\overline{y}_{PA} = 50,000$ and the sample SD, $\sigma_{\overline{y}} = 5,000$, $N = 2,000$**
- **What's the likelihood we would observe 50,000 in the sample if $\mu_{PA} = 70,000$?**

Distribution of Repeated Samples calculating $\overline{y}_{PA}$



$\mu_{PA} = 70{,}000$

Actually Very Easy to Answer this Now

- We need to calculate the z-score for 50,000 if the null hypothesis is true: $z = \frac{\bar{x}-\mu_x}{\frac{\sigma_{\bar{x}}}{\sqrt{N}}} =$
  $\frac{50{,}000-70{,}000}{\frac{5{,}000}{\sqrt{2{,}000}}} = \frac{-20{,}000}{111.86} = -178.79$
- Now, look at the Z table to see how likely it would be that we see this sample mean if the null hypothesis is true

Standard Normal Probabilities

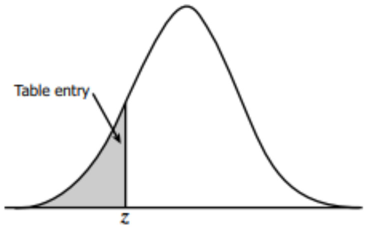


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

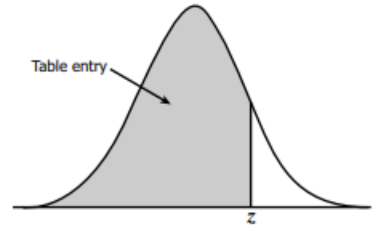


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Probability is basically 0. "We reject the null hypothesis above 99% confidence."

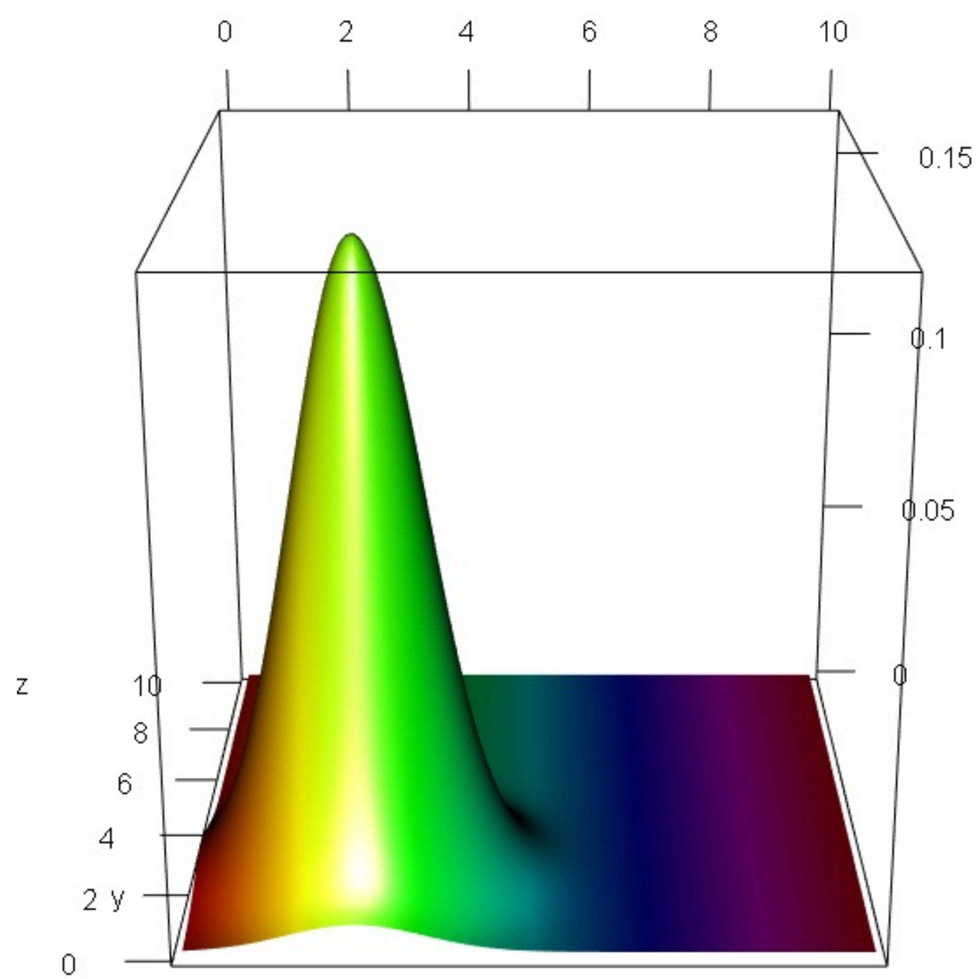
Student's *t* Distribution

- **In practice we'd use Student's t distribution instead of the standard normal distribution in these calculations**
- **Student's t distribution is a distribution related to the standard normal distribution**
  - We'll talk about it more later, but for now
  - Think of it as a distribution very similar to the normal distribution that creates more uncertainty (larger variance) when certain information is lacking
- **Originally used when sample sizes were small**
- **Practice has evolved to use it whenever approximating a population variance with the sample variance**
- **We'll be coming back to Student's t distribution later**

Joint Probability Density Functions

- **Of course in econometrics we want to know about <u>relationships</u> between two or more variables**
  - So we need to be able to think about probabilities two values are observed together
- **The <u>joint probability density function</u> describes this**
  - For two or more random variables a JPDF describes all the possible combinations of values that the set of random variables might take on and probabilities with which the set of random variables takes on specific values or values in specific ranges

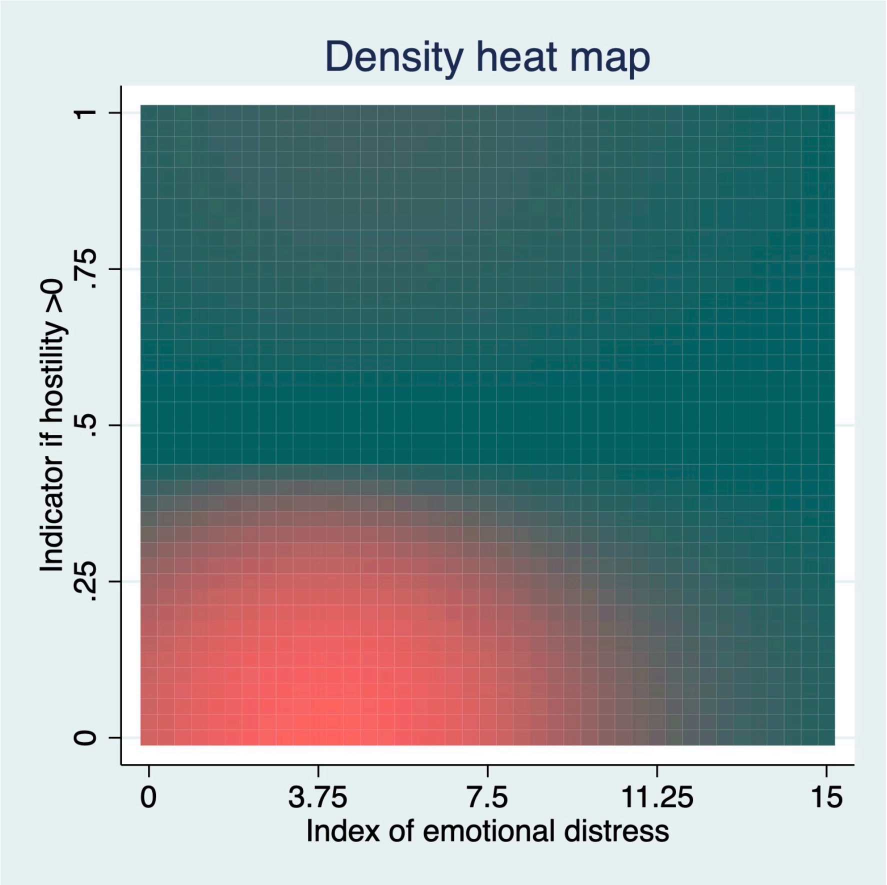
Joint Probability Density Functions

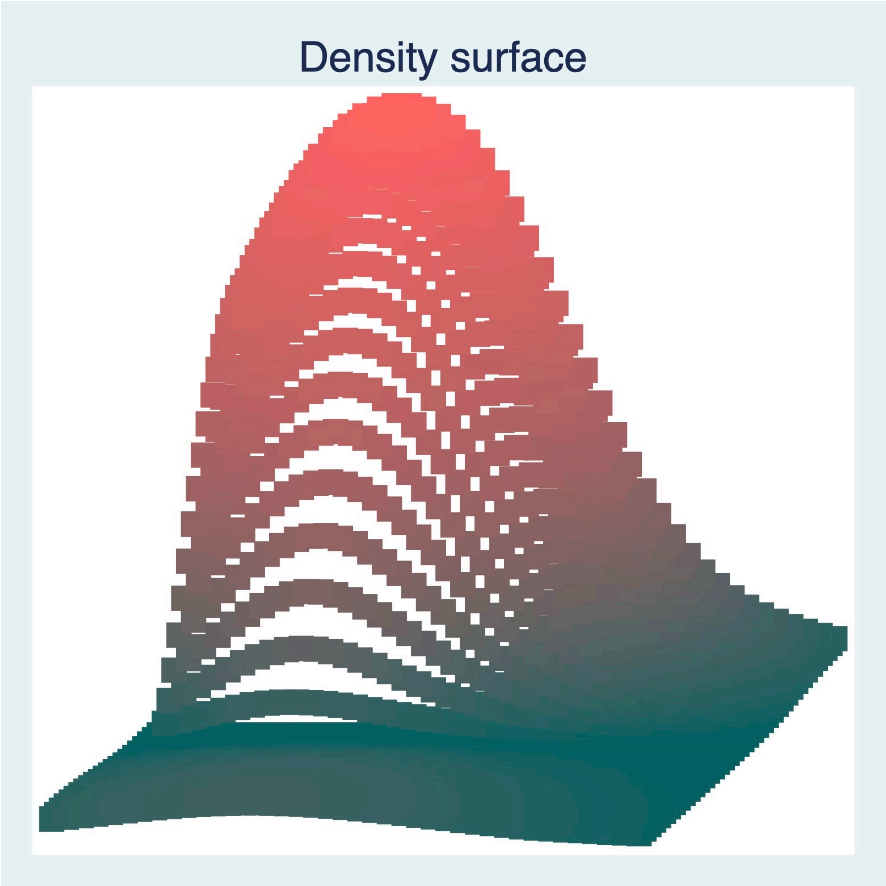


- You can think of a normal joint PDF as a "mountain" rising into a third dimension from a two dimensional diagram

Joint Probability Density Functions

- **You can also display joint PDFs in the form of a heatmap**
  - These are produced using the `tddens` command in Stata





Measures of Covariation

- **With joint PDFs, we're interested in summarizing certain features of the distribution:**
  - The means of all the jointly distributed random variables
  - The variance of all the jointly distributed random variables
  - **A measure of the extent to which various pairs of random variables "move together"**
    - If one variable has a high value, does the other one have a high value too or a low one?
    - How do we measure this?

Measures of Covariation

- **The primary measure we use is "covariance" between Y_1 and Y_2, denoted $Cov(Y_1, Y_2)$ or σ_{Y_1,Y_2},**

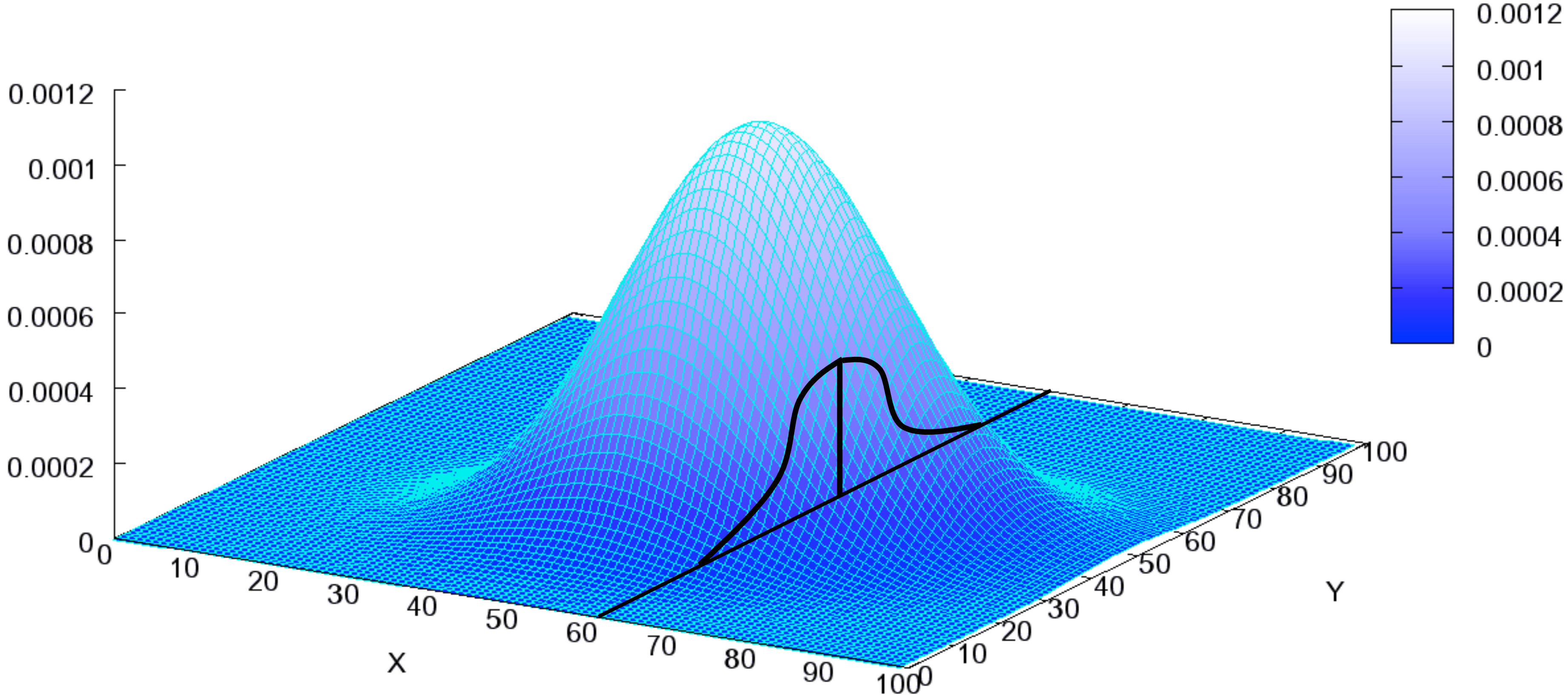
$$Cov(Y_1, Y_2) = \frac{\sum_{i=1}^{N} (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{N} \qquad (11)$$

Conditional Distributions

- **When two random variables have some relationship with each other, knowing the value of one of them influences our thoughts about the likely values of the second**
  - Suppose two random variables both have a mean of zero and they are positively correlated
    - If we know that one observation had a high positive value on a particular trial, then our best guess for the value of the second variable would be high and positive as well
- **A "conditional distribution" describes our thoughts about the likely values of a random variable given what we know about other random variables**

Visualizing Conditional Distributions

The Conditional PDF can be pictured by looking at the variation in the Y values, once you know a value, or range of values that you're interested in X



Using Conditional Distributions in Econometrics

- The models we estimate in econometrics describe how conditional expectations of certain random variables are related to the values of other random variables on which we condition
- For example, we might be interested in the following relationship,

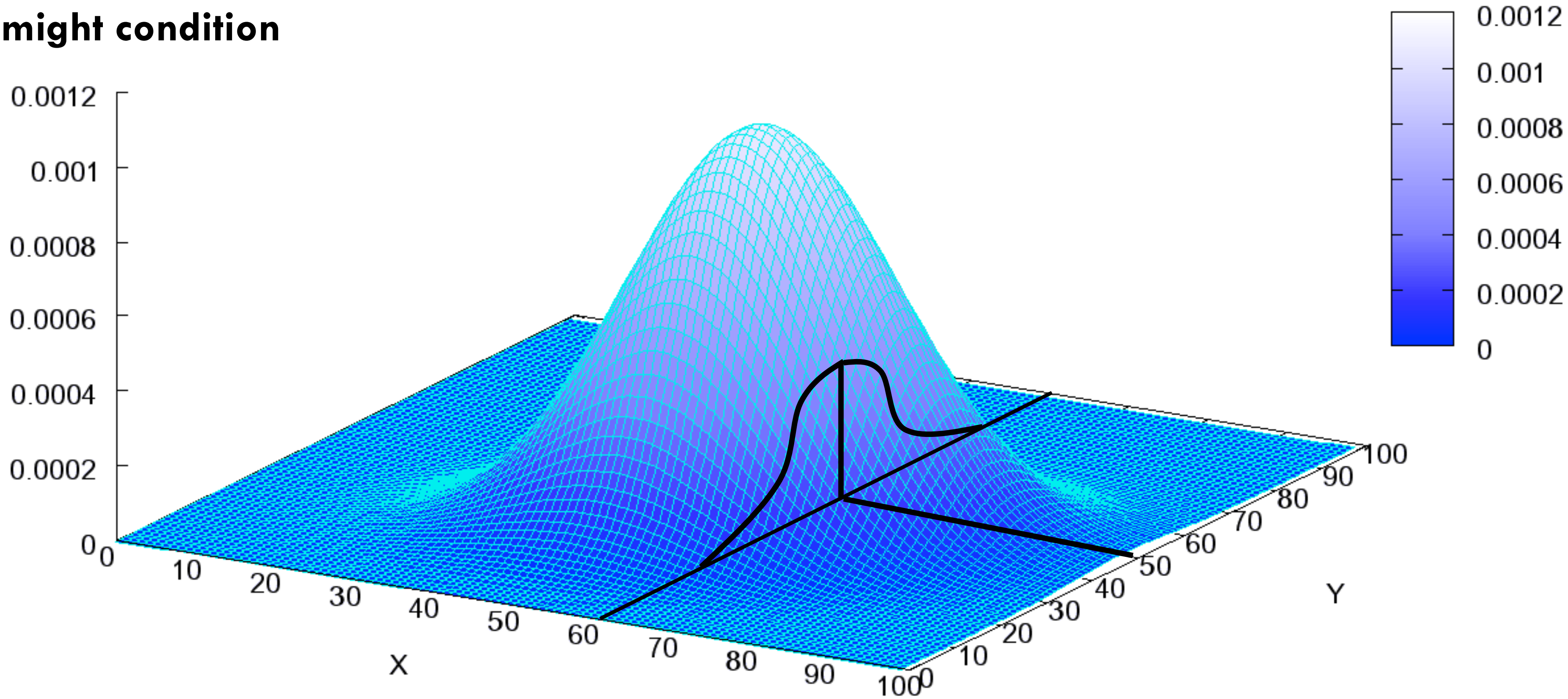
$$E(Y|X) = \beta_0 + \beta_1 X \tag{14}$$

- Both Y and X are random variables, <u>but we're interested in understanding how the conditional expectation of Y varies as we vary the level of X</u>

Visualizing Conditional Distributions

$E(Y|X = 60) = 50$ in this picture

Our econometric models will describe the expected value of Y for every value of X (and other variables in more elaborate models) on which we might condition



Useful Properties of Expectation Algebra and Variances of Linear Functions of Random Variables

- **Let a, b, and c be scalars (constants). Let X and Y be random variables**

$$E(a) = a \tag{15}$$

$$E(bY) = bE(Y) \tag{16}$$

$$E(a + bY) = E(a) + E(bY) = a + b(Y) \tag{17}$$

$$E(X + Y) = E(X) + E(Y) \tag{18}$$

Useful Properties of Expectation Algebra and Variances of Linear Functions of Random Variables

- **Let $a()$, $b()$, and $c()$ be functions. Let X and Y be random variables**
  - Read "$E(Y|X)$" as "the expected value of Y conditional on X taking some particular value"
- **If X and Y are independent $(X \perp Y)$, then $E(Y|X) = E(Y)$**
- $E(c(X)|X) = c(X)$

Stochastic Independence

- **Random variables in some set are "independent" or "stochastically independent" if knowing the value that any one of them takes on has no influence on our expectations about the others**
  - In other words, the conditional distribution of each random variable in the set is the same, no matter what values of the other random variables you condition on
- **Examples:**
  - Roll of a die after the roll of another die
  - Pulling a name out of a hat with replacement

Stochastic Independence

- **If two random variables X and Y are independent, then $Cov(X,Y) = 0$**
  - The reverse is not necessarily true (although it is if X and Y are normally distributed)