Distributions Overview

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

February 5, 2014

Outline

This is The Lecture That Accompanies My essay, "Distribution Overview: Probability by the Seat of the Pants"

1. What is Probability?
2. Characterizing Distributions
   - Expected Value
   - Variance
3. Algebra of Expected Values and Variances
4. Example Distributions
   - Exponential
   - Normal
   - Gamma
   - Beta
   - χ^2 (Chi-Squared)
   - t
   - F
   - Binomial
   - Poisson
5. Practice Problems

Take Away Points to Watch for

- Key Terms:
  - probability distribution
  - random variable
- Characterizing Distributions
  - Parameters are "knobs"
  - More-or-less universally applicable characteristics of distributions
    - Expected Value
    - Variance

Outline

1 What is Probability?
2 Characterizing Distributions
- Expected Value
- Variance

3 Algebra of Expected Values and Variances
4 Example Distributions
- Exponential
- Normal
- Gamma
- Beta
- χ^2 (Chi-Squared)
- t
- F
- Binomial
- Poisson

5 Practice Problems

Probability: Outcomes and Chances

- list a set of possible outcomes, $X = \{x_1, x_2, x_3, \ldots\}$. A "Sample Space" from which observations are drawn.
- Specify the probability of each one.

$$p(x_i) \geq 0 \tag{1}$$

"Probability Mass Function" if X is a discrete set

- the sum is unity

$$\sum_{i=1}^{m} p(x_i) = 1.0 \tag{2}$$

- If it does not add up to 1.0, we can always force it by dividing by the sum. This "normalizes" it, replacing probabilities by $p(x_i)/"sum"$

Simple Example

- Consider a discrete variable that can take on values $\{1, 2, 3, 4, 5, 6\}$.
- A probability model must calculate the $p(x_i)$ for each one.

Outcome	$x_i =$	1	2	3
probability	$p(x_i)$	1/6	1/6	1/6

Outcome	$x_i =$	4	5	6
probability	$p(x_i)$	1/6	1/6	1/6

- Here we suppose the 6 outcomes are equally likely.

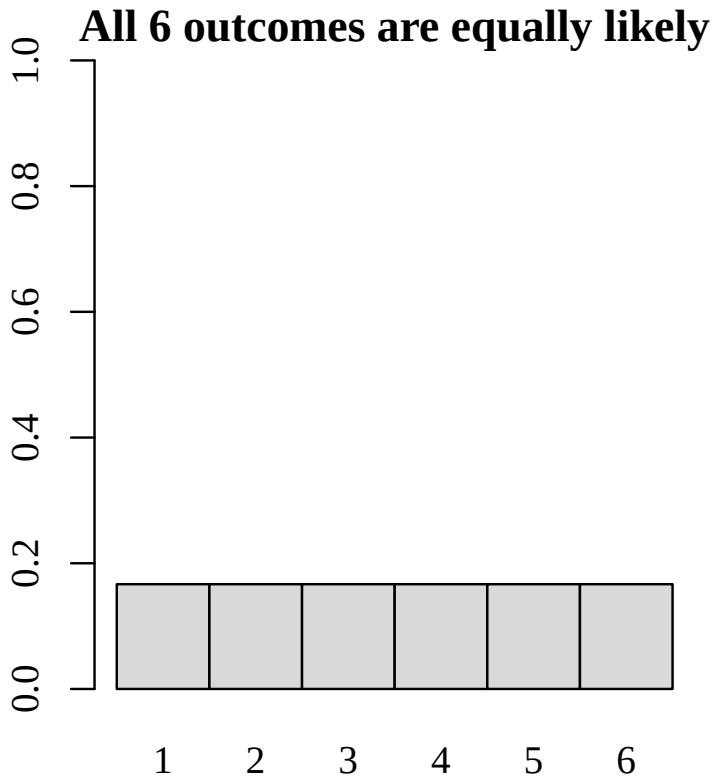


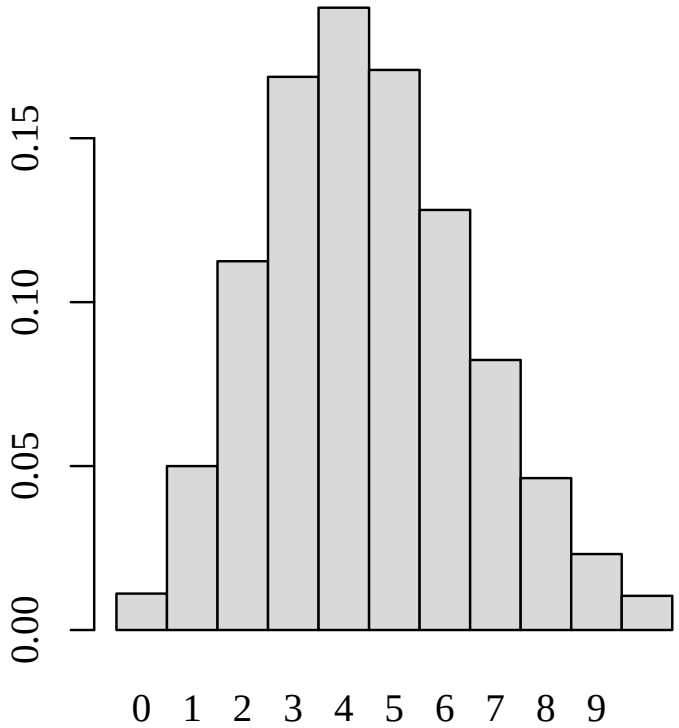
Illustration 2: More Variety

- Stair step illustration
- $X = \{0, 1, 2, \ldots, 10\}$

x_i	0	1	2
$p(x_i)$	0.01	0.05	0.11

3	4	5	6
0.17	0.19	0.17	0.13

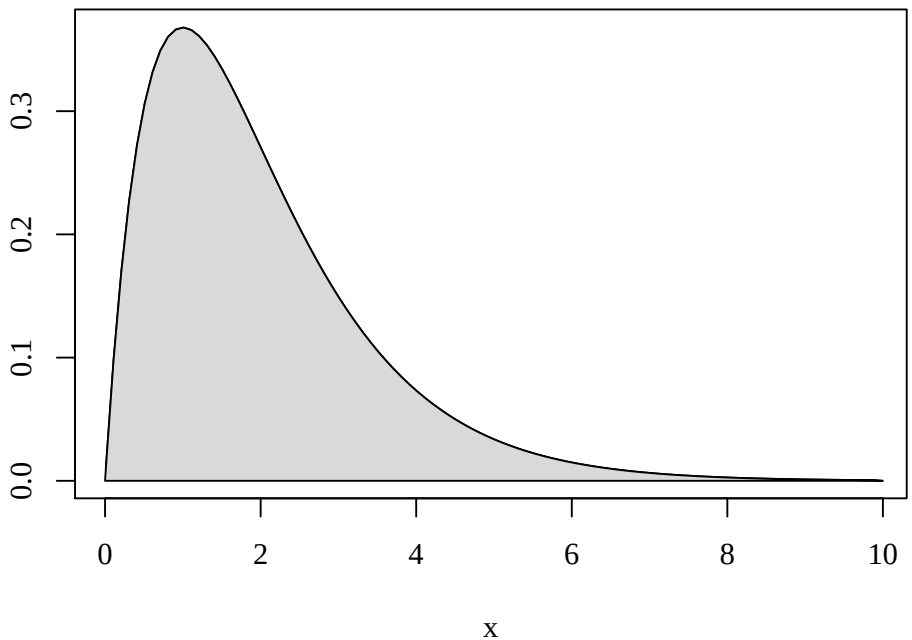
7	8	9	10
0.08	0.05	0.02	0.01



If X is a continuum

- If X is a subset of the real number line, $\mathbb{R}$, we can't list all outcomes one by one.
- "Probability Density Function"(PDF)
  - $f(x) \geq 0$ and
  - $\int_X f(x)\,dx = 1$

  $\int_X f(x)\,dx$ means we are collecting the area above X.



Anything Can Become a PDF (almost)

- Take any function $f(x)$
- Suppose the area under $f(x)$ between a and b is defined (can be calculated).
- A PDF is created if we divide $f(x)$/"*area under* $f(x)$".
- Partly for this reason, we have too many distributions to study (Compendium of distributions includes at least 150 distributions).

What Do Probability Numbers Really Mean?

- Consider the probability of one value from a random process. AKA "an observation" or "a variate" or "a sample"
- The chance of an orange is 0.3 at lunch. The chance of rain before midnight is 0.20.
- What does the probability number mean? (deep philosophical problem)
- interpretation 1: "long run relative frequency". Take an infinite number draws. The probability of x_i is the fraction

$$lim_{\#of\ draws \to \infty} \frac{\#observed\ x_i}{\#of\ draws} \tag{3}$$

- Interpretation 2: "Bayesian" relative likelihood interpretation. Probability is the "degree of belief" one draw will yield a particular result.

Additive Property

- Dice: Playing craps? You may need to know the chance of 6 or 3
- Betting on Baseball? The chance of a hit is the chance of a single + chance of double + chance of triple + chance of homer
- We ADD together the chances when considering the chance of different outcomes occurring in a single draw.

Theorem

"or" is Addition: Consider 1 "trial". The chance that either x_i or x_j will happen equals the sum of the chance that each one individually will happen.

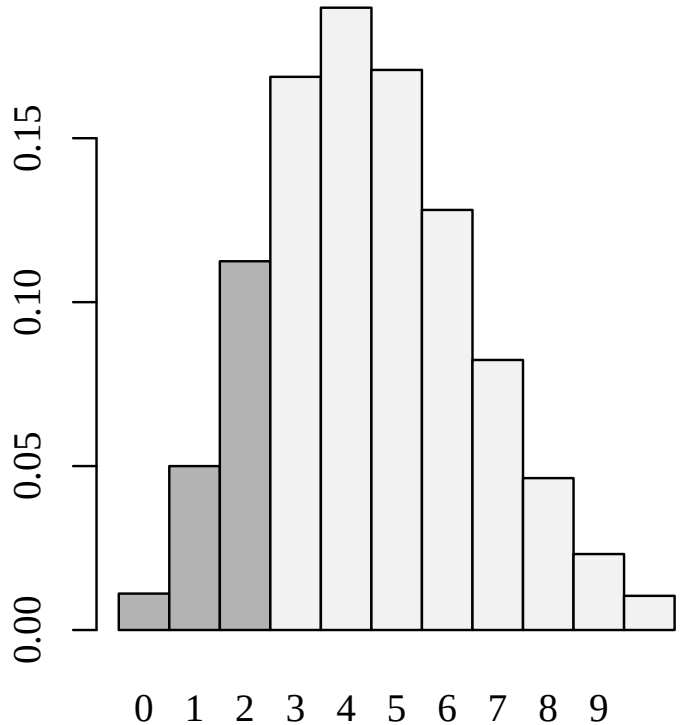
$$probability(x_i \text{ or } x_j) = p(x_i) + p(x_j) \tag{4}$$

Illustration: Addition Property with Discrete Distribution

- Chance of outcome smaller than 3 is Sum

$$\sum_{x<3} p(x)$$

- $p(0) + p(1) + p(2)$



Multiplication Property

Consider 2 independent "trials" or "samples" from a given probability process x_i and x_j.

Theorem

"and" is Multiplication: The chance that both x_i and x_j will happen is the product

$$probability(x_i \text{ and } x_j) = p(x_i) \times p(x_j) \tag{5}$$

Here the sample space includes 2-tuples, (x_i, x_j). The notation for the space of all possible pairs is $X \times X$, or X^2. That's called a "Cartesian product".

Multiplication and Sampling

- We use the multiplication Principle ALL THE TIME in the analysis of samples.
- We assert observations in a sample $\{y_1, y_2, \ldots, y_N\}$ are statistically independent
- The probability of observing the "whole sample" is the product of the individual probabilities

$$p(y_1, y_2, y_3, \ldots, y_N) = p(y_1) \cdot p(y_2) \cdot p(y_3) \cdots p(y_N) \tag{6}$$

- We often try to decide if a probability theory is correct by figuring out how likely a sample is (Maximum Likelihood Analysis)

Illustration: Multiplication Property with Discrete Distribution

- Draw one score, y_1
- Draw another score, y_2
- What is the chance that both y_1 and y_2 are equal to 3?
  $p(y_1 = 3) \cdot p(y_2 = 3)$

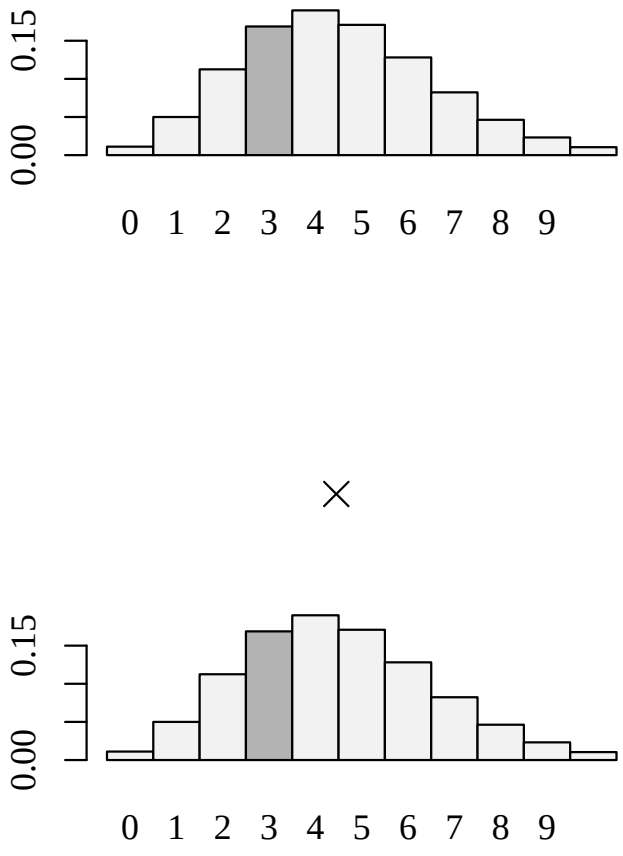
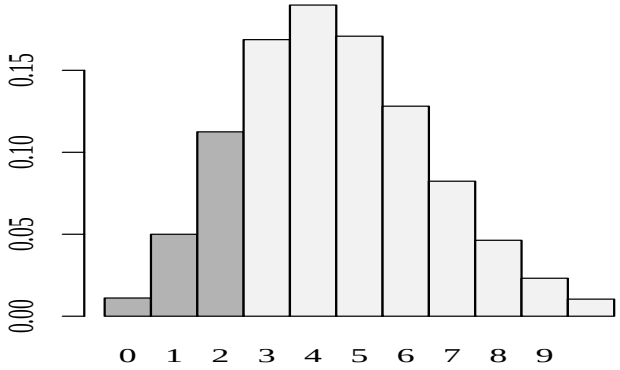


Illustration: Multiplication Property with Discrete Distribution

- What is the chance that both y_1 and y_2 are smaller than 3?

$$p(y_1 < 3) \cdot p(y_2 < 3) =$$
$$(p(y_1 = 0) + p(y_1 = 1) + p(y_1 = 2)) \times$$
$$(p(y_2 = 0) + p(y_2 = 1) + p(y_2 = 2))$$



$\times$

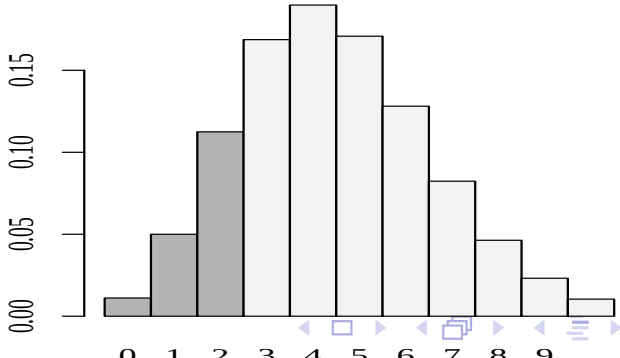
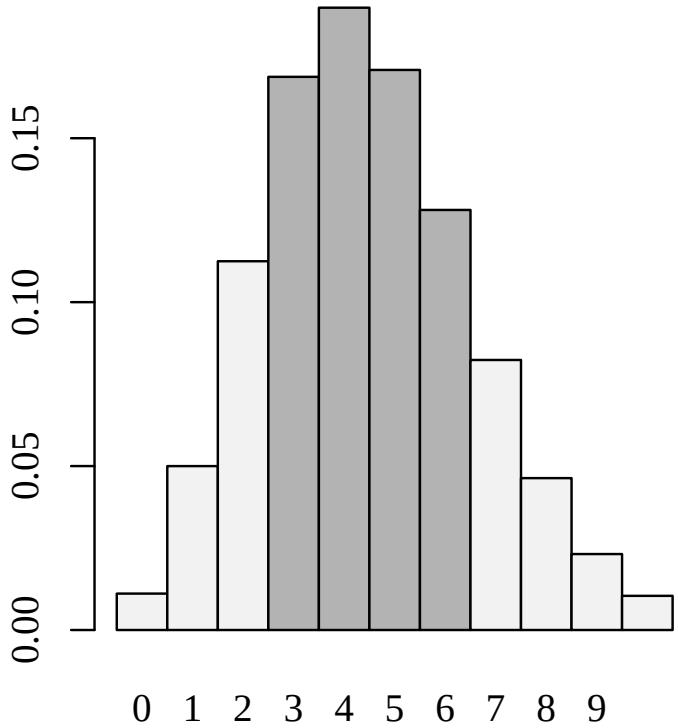


Illustration: Discrete Distribution

- Chance that $y_1 = 4$ and $3 \leq y_2 < 7$?
- Duh! $p(y_1 = 4) \times p(3 \leq y_2 < 7)$
- Note chance of a "slice" is difference between cumulative values:
  $p(3 \leq y_2 < 7) = p(y_2 < 7) - p(y_2 < 3)$
  So:
  $p(y_1 = 4) \times \{(p(y_2 < 7) - p(y_2 < 3)\}$

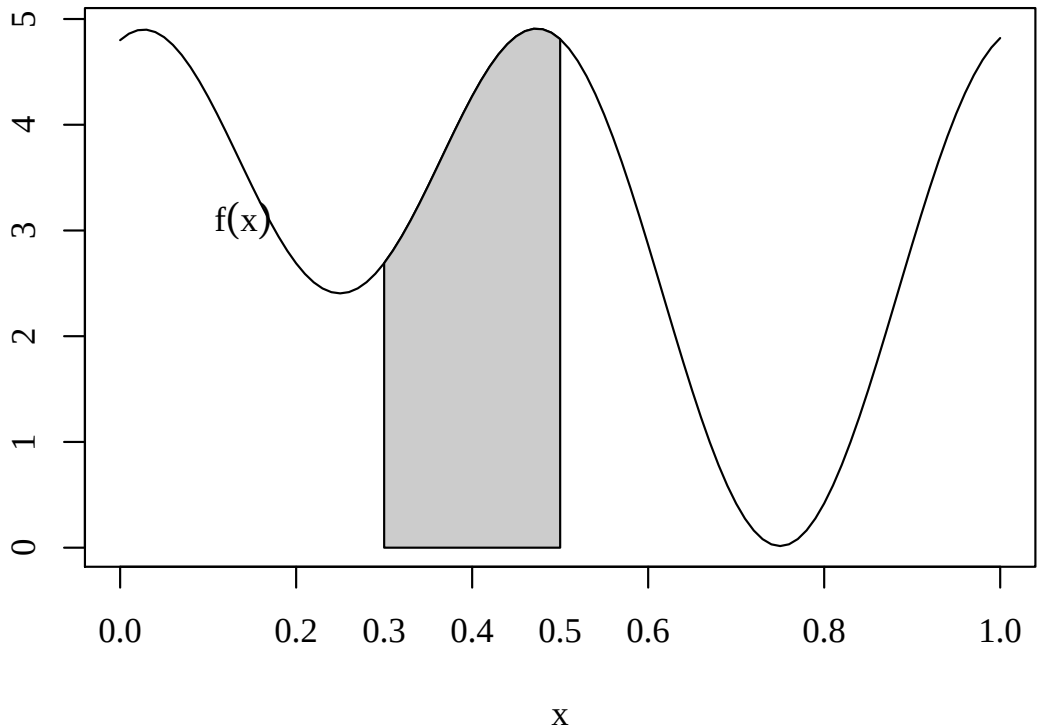


Discrete Versus Continuous Jargon

- Discrete distributions: PMF, $p(x_i; \textit{some parameters})$ "probability mass function"
- Continuous distributions:
  - PDF, $f(x; \textit{some parameters})$ "probability density function".
  - CDF, $F(x^u; \textit{parameters})$ "cumulative distribution function", chance that random draw will be smaller than x^u.
  - PDF problem: chance of one point occurring is 0 because each point has "no area above it" problem.
  - CDF allows us to discuss outcomes in regions between 2 points $k1$, $k2$

$$F(k_2) - F(k_1) \tag{7}$$

Illustration: Continuous Distribution



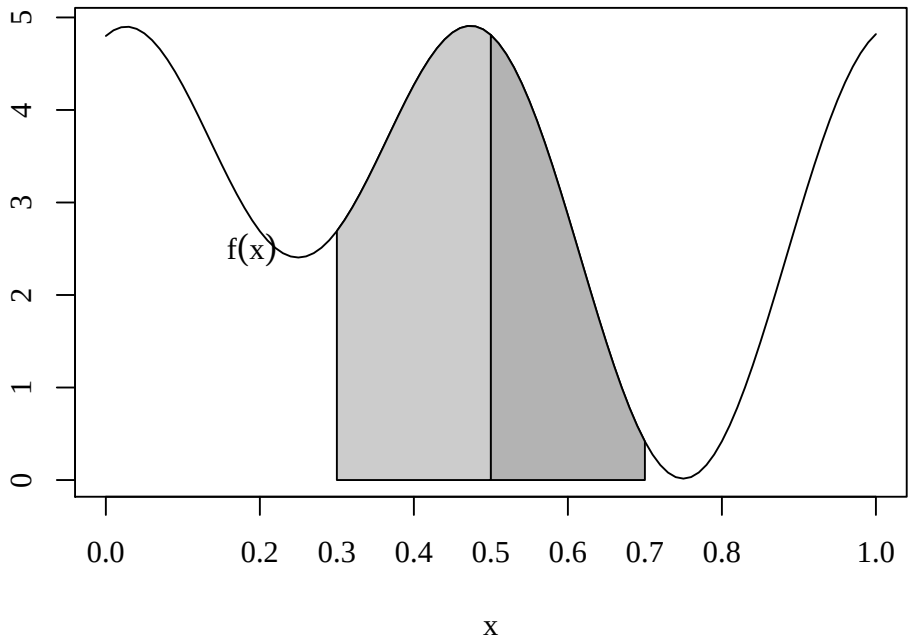
Shaded area is probability that one observation will be between 0.3 and 0.5=

$$\int_{0.3}^{0.5} f(x)\,dx = F(0.5) - F(0.3)$$

Illustration: Addition with a Continuous Distribution

- The chance of an outcome between 0.3 and 0.7?
  $F(0.7) - F(0.3)$
- Area between 0.3 and 0.7= sum of area between 0.3 and 0.5 and 0.5 and 0.7.
- Obviously,

$$Pr(0.3 < y_i < 0.7) =$$
$$F(0.7) - F(0.3) =$$
$$F(0.5) - F(0.3) + F(0.7) - F(0.5)$$



Outline

1 What is Probability?

2 Characterizing Distributions
- Expected Value
- Variance

3 Algebra of Expected Values and Variances

4 Example Distributions
- Exponential
- Normal
- Gamma
- Beta
- χ^2 (Chi-Squared)
- t
- F
- Binomial
- Poisson

5 Practice Problems

Distributions: Theories

- Random variables result from a "data generating process" = "stochastic process"
- Need ways to describe them.
- Sometimes this is called a "population", in the sense it is the thing from which observations are drawn.
- Predict obvious confusion, where students think "population" is population in the vernacular meaning

Expected Value

Definition: probability weighted sum of outcomes. $\sum probability \times outcome$

discrete

$E[x] = \sum_{i=1}^{m} p(x_i) \cdot x_i$

This pre-supposes the outcomes are numeric, like dice or other numbered things.
Dice: $\frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6$

continuous

$E[x] = \int_a^b f(x) \cdot x \, dx$
where x is defined on (a, b)

Tempting to say: expected value is the "mean of population of x" but I find that esoteric, since we don't agree on what population means

Why a big deal? Even a single observation is characterized by variance and expected value

Terminology: "Expected". Really?

- Not "single most likely outcome," it is not what I "expect" (subjectively)
- Not necessarily "in the middle of the distribution".
- A sample mean is an estimate of the Expected Value.

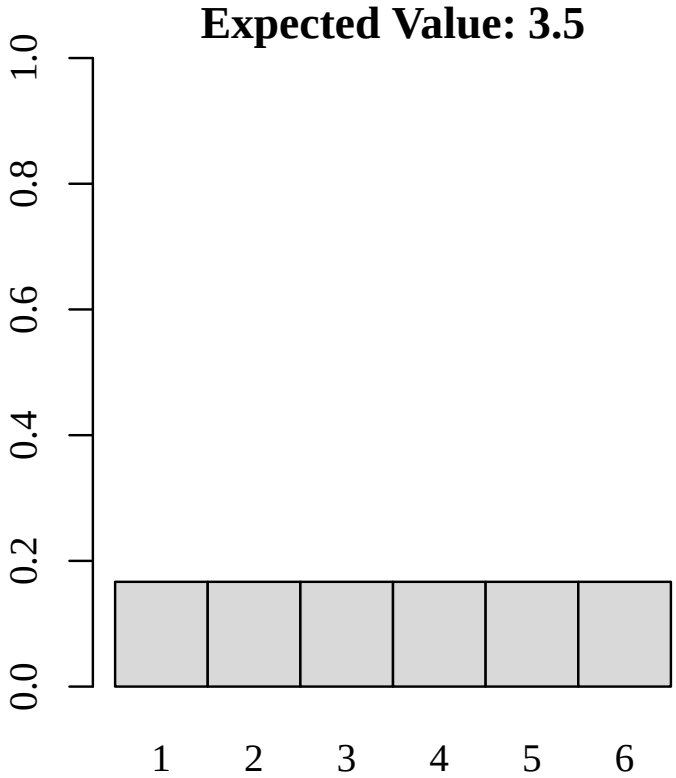
$E[x]$ More Obviously Meaningful Sometimes

Claim

If PDF is
1) symmetric and 2) unimodal, then
E[x]== Mode[x] == Median[x].

- So, for some distributions, the Expected Value does have a visual "handle" we can hold on to.
- However, for just as many distributions, the number we get when calculating the "expected value" is not a number I would have "expected" (subjectively speaking)

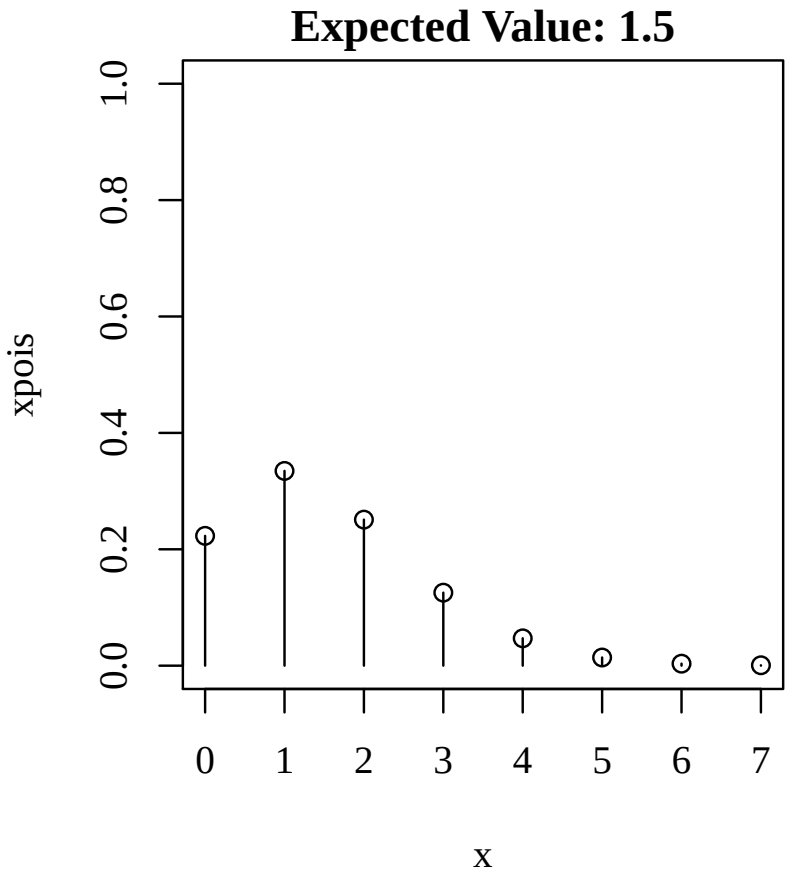
Unimodal Example: Expected Value is the Median (&mode)



What's neat about that?

- Outcomes are discrete 1-6, the EV is real number 3.5
- "Expected" is the officially accepted term here, but 3.5 is not "expected" subjectively

Counter Example: Expected Value is 1.5



I used spikes rather than bars here, to emphasize discreteness of x

- Does it help you to know EV is 1.5?

How much variety is possible?

- If the probability of all outcomes is near 0, except for 1, there's not much variety.
- If all outcomes are equally likely (uniform distribution), there's maximum possible variety.
- We need a way to summarize the "in between" cases.
- In descriptions of samples, we used variance, $\frac{1}{N}\sum(x_i - \bar{x})^2$, the mean of squared deviations.
- To describe probability models, we take very similar approach, also using term Variance

Var[x]: Variance of a Random Variable

Definition: expected value of $(x - E[x])^2$. Expected squared deviations...

- Repeat: Variance is the expected squared deviation about the expected value

$$Var[x] = \sum p(x_i) \cdot (x_i - E[x])^2 \tag{8}$$

- intuition: if values of x_i are spread out "far and wide," then $Var[x]$ will be a bigger number

The Standard Deviation of a Random Variable

- Definition: Std. Deviation of a random variable is the square root of its variance.
- The *Standard Deviation* scales proportionally!

$$Std.Dev.[k \cdot x] = k \cdot Std.Dev.[x] \tag{9}$$

- Thus, the ratio of the expected value and standard deviation is not affected by k.

$$\frac{E[x]}{Std.Dev.[x]} = \frac{E[k \cdot x]}{Std.Dev.[k \cdot x]} = \frac{k \cdot E[x]}{k \cdot Std.Dev.[x]} = \tag{10}$$

Difference between Sample mean and E.V.

- The mean (or "average") of a sample is an estimate of $E[x]$
- The "average" varies from sample to another
- But the $E[x]$ is the same, it is a characteristic of the data generating process.

Notation: μ_x ,$\bar{x}$, $E[x]$, $\widehat{\mu_x}$, $\widehat{E[x]}$

- $E[x]$ sometimes referred to as "mu", the Greek μ_x. It is a theoretical quantity, NOT an estimate from a sample
- Notation for sample estimates:
  - $\bar{x}$ is very widely used.
  - I'd like to call an estimate of that $\widehat{E[x]}$ or $\widehat{\mu_x}$, just to keep notation simpler

Algebra of Variance: The standard deviation of $\bar{x}$

- The variance of the sampling distribution shrinks as the sample size is increased.
- There's a section below on "Algebra of Expectations and Covariance" that develops the following result formally, but here it is, just for fun.
- Suppose we repeatedly draw samples of size N from a random process. The variance of the sampling distribution of $\bar{x}$ (across repeated samples) is

$$Var[\bar{x}] = \frac{1}{N} Var[x] \tag{11}$$

Here's the proof.

- Let the variance of x be $Var[x]$.
- The mean of x is

Algebra of Variance: The standard deviation of $\bar{x}$ ...

$$\bar{x} = \frac{1}{N}x_1 + \frac{1}{N}x_2 + \ldots + \frac{1}{N}x_N \tag{12}$$

Apply the $Var[]$ function to both sides:

$$Var[\bar{x}] = Var[\frac{1}{N}x_1 + \frac{1}{N}x_2 + \ldots + \frac{1}{N}x_N] \tag{13}$$

To simplify, assume $Cov[x_i, x_j] = 0$, so:

$$Var[\bar{x}] = \frac{1}{N^2}Var[x_1] + \frac{1}{N^2}Var[x_2] + \ldots + \frac{1}{N^2}Var[x_N] \tag{14}$$

All of the $x's$ are from the same random process, so their variances are all the same, $Var[x]$.

$$\begin{aligned} Var[\bar{x}] &= \frac{N}{N^2}Var[x] \\ &= \frac{1}{N}Var[x] \end{aligned} \tag{15}$$

Algebra of Variance: The standard deviation of $\bar{x}$ ...

- This is an ESSENTIAL component in the process of inferential statistics. We have an avenue from the observation of x's variance within a sample to a view of $\bar{x}$'s variance across many samples.
- Standard deviation of the average, known as the "Standard error of the mean of x",

$$Std.Err(\bar{x}) = Std.Dev[\bar{x}] = \frac{1}{\sqrt{N}} Std.Dev.[x] \tag{16}$$

Outline

1 What is Probability?
2 Characterizing Distributions
- Expected Value
- Variance

3 Algebra of Expected Values and Variances
4 Example Distributions
- Exponential
- Normal
- Gamma
- Beta
- χ^2 (Chi-Squared)
- t
- F
- Binomial
- Poisson

5 Practice Problems

Algebra of Expected Values

Proportional Scaling. If x is re-scaled $k \times x$, the Expected Value of $k\,x$ is easy to calculate. Multiply k times the original $E[x]$

- $E[k \cdot x] = k \cdot E[x]$, where k is a "constant"

So, for example, if I said the expected value of variable *fish* is 10. Some GRA re-scales the fish variable by dividing by 10. The new variable $newfish = 0.1 \times fish$, the expected value of *newfish* is 1.

- Now consider 2 random variables. $x1$ and $x2$.

1. Additivity. The expected value of a sum is the sum of the expected values
   - $E[x1 + x2] = E[x1] + E[x2]$
2. Linearity. Combine the Scaling and Additivity

- $E[k_1x1 + k_2x2] = k_1E[x1] + k_2E[x2]$

New Term: Covariance

- Covariance: How much do 2 random variables "go together"
- Are they both above their expected values at the same time? Or both below?
- The expected value of the product $(x1 - E[x1]) \times (x2 - E[x2])$
- Write it out

$$Cov[x1, x2] = E[(x1 - E[x1]) \cdot (x2 - E[x2])] \tag{17}$$

- Use of discrete variables, we can write down a sum.

$$Cov[x1, x2] = \sum p(x1_i, x2_i) \cdot (x1_i - E[x1])(x2_i - E[x2]) \tag{18}$$

- I wrote a little R script to help visualize covariance. It should be in this folder, "distro-covar-1.R"

Algebra of Variance

- Important fact 1: The variance of $(k \cdot x)$ is k^2 times the variance of x.

$$Var[k \cdot x] = k^2 \cdot Var[x] \tag{19}$$

- Interesting Tidbit: The variance of x is equal to the expected value of x-squared ($E[x^2]$) minus the square of the expected value of x ($(E[x])^2$)

$$Var[x] = E[x^2] - (E[x])^2 \tag{20}$$

Proof.

$$Var[x] = E[(x - E[x])^2] = E[x^2 - 2E[x] \cdot x + (E[x])^2]$$

$$= E[x^2] - (E[x])^2$$

We'll see many applications of this basic idea later on □

Variance of a Sum

- Add two random variables, $x1$ and $x2$, their variance combines 2 variances plus 2 times their covariance:

$$Var[x1 + x2] = Var[x1] + Var[x2] + 2Cov[x1, x2] \tag{21}$$

- If $x1$ and $x2$ have weights, we can carry them through

$$Var[k_1x1 + k_2x2] = k_1^2 Var[x1] + k_2^2 Var[x2] + 2k_1k_2 Cov[x1, x2] \tag{22}$$

- $2k_1k_2 Cov[x1, x2]$ is a hassle that we run into all the time.
- We'd like some reason, any reason, to assume it away, so we can simplify that:

$$Var[k_1x1 + k_2x2] = k_1^2 Var[x1] + k_2^2 Var[x2] \tag{23}$$

Outline

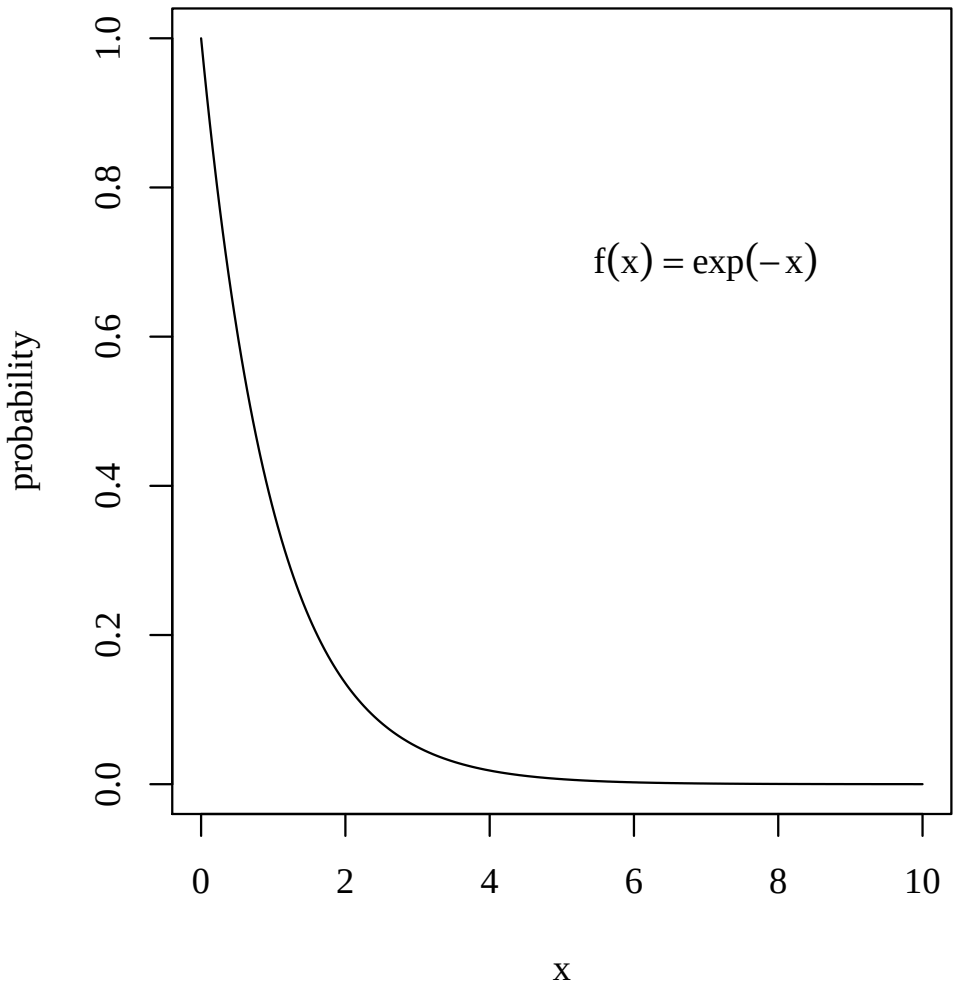
1 What is Probability?
2 Characterizing Distributions
- Expected Value
- Variance

3 Algebra of Expected Values and Variances
4 Example Distributions
- Exponential
- Normal
- Gamma
- Beta
- χ^2 (Chi-Squared)
- t
- F
- Binomial
- Poisson

5 Practice Problems

Exponential Distribution

- Time that one must wait before an "event" occurs if the chance of an event depends only on the amount of time that passes.
- If the probability of an "event" is $\lambda \cdot \Delta t$ (for Δt shrinking to 0), then the time waited before an event is exponentially distributed.



Probability Density Function

$$f(x;\lambda) = \lambda e^{(-\lambda x)}, \text{ where } x \geq 0 \tag{24}$$

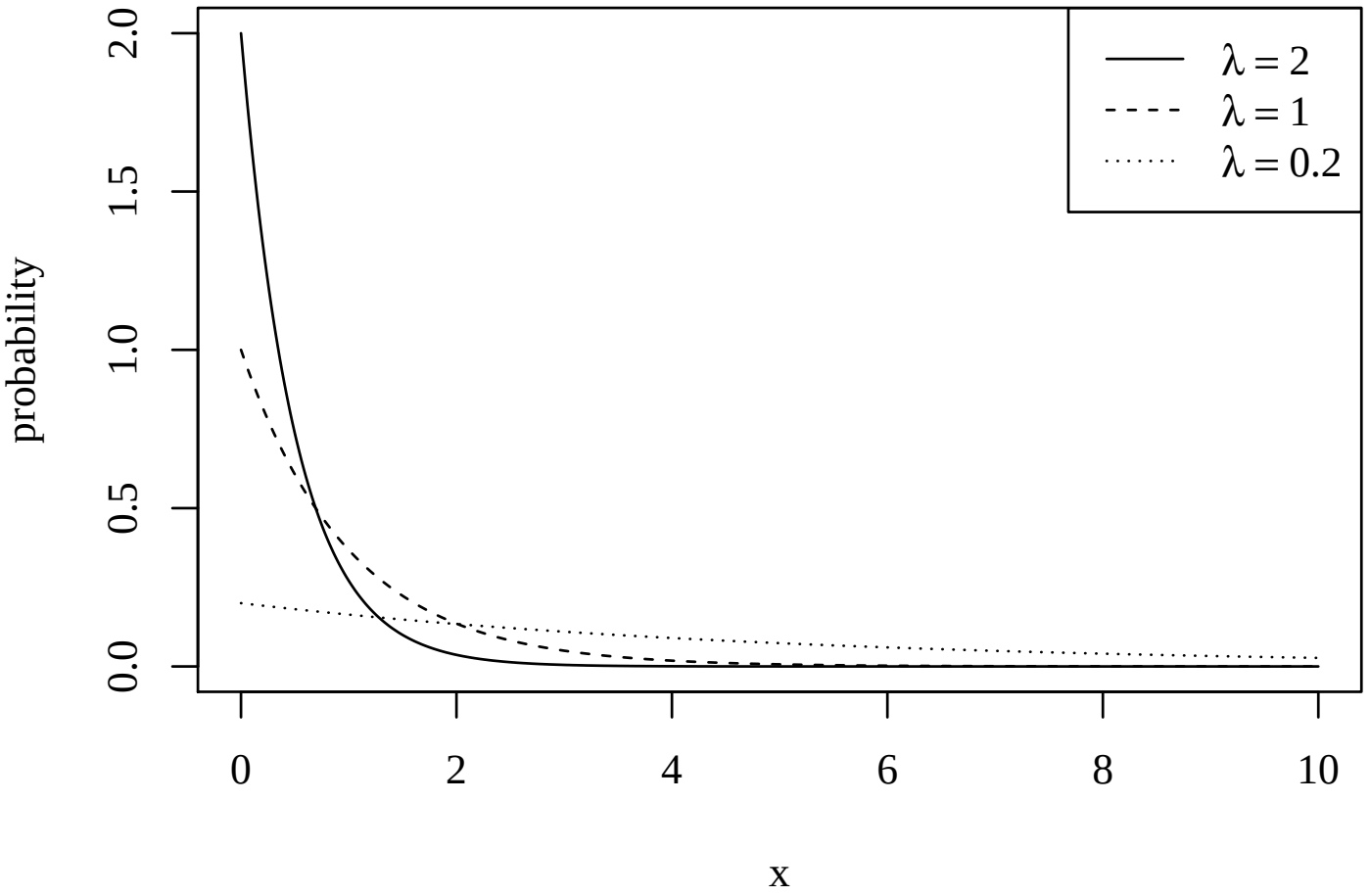
λ, which is called the "rate" parameter.
Some books use the reciprocal of λ, so the density would be

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu}$$

- Other notations:
  - $f_\lambda(x)$.
  - $f(x)$ parameters are implicit.

More

If λ is very small, the decline in the value of $f(x;\lambda)$ is very gradual.



Cumulative Distribution Function

$$
\begin{aligned}
F(k;\lambda) &= \int_0^k \lambda e^{-\lambda x} dx \\
&= -e^{-\lambda x} \Big|_0^k \\
&= 1 - e^{-\lambda k}
\end{aligned} \tag{25}
$$

Moments

$$E[x] = \frac{1}{\lambda} \tag{26}$$

$$Var[x] = \frac{1}{\lambda^2} \tag{27}$$

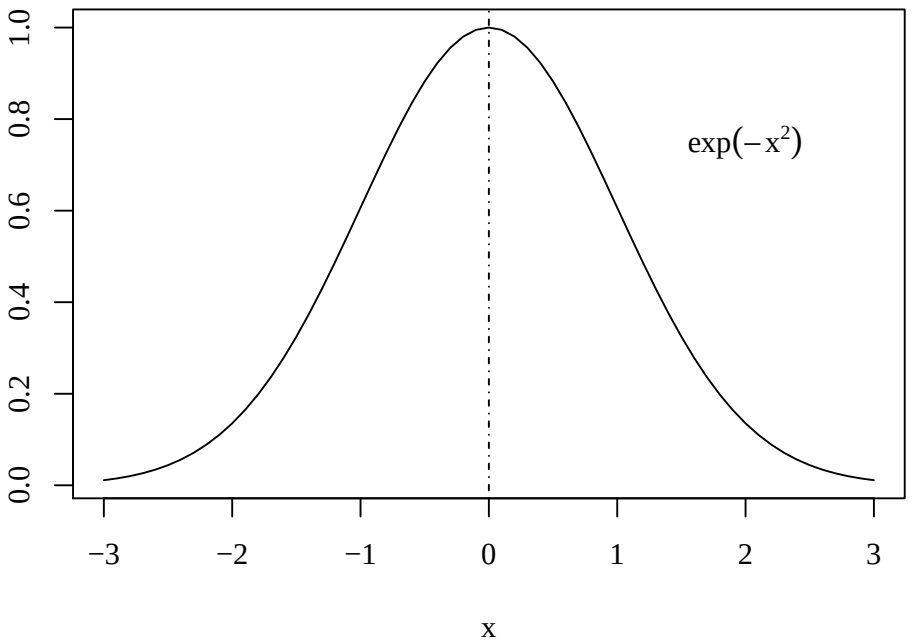
Normal Distribution

- PDF:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \tag{28}$$

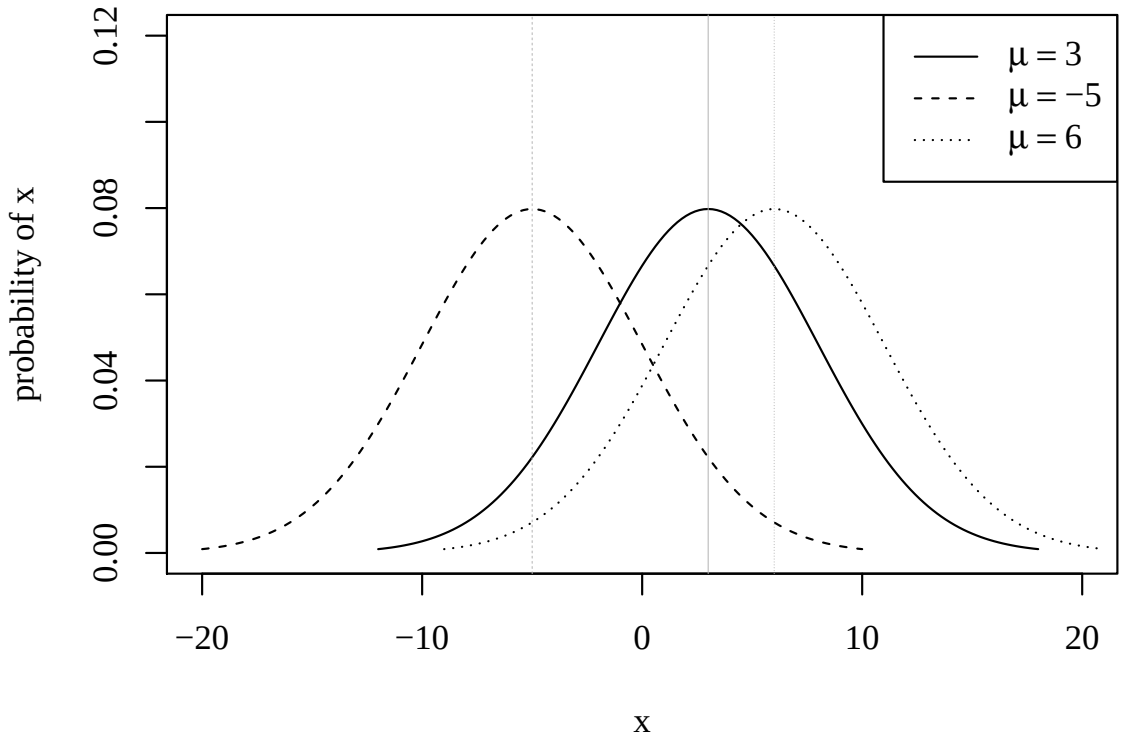
The normalizing constant, $1/\sqrt{2\pi\sigma^2}$.

- Essence of it is $exp(-x^2)$.
- Uni-modal and symmetric.



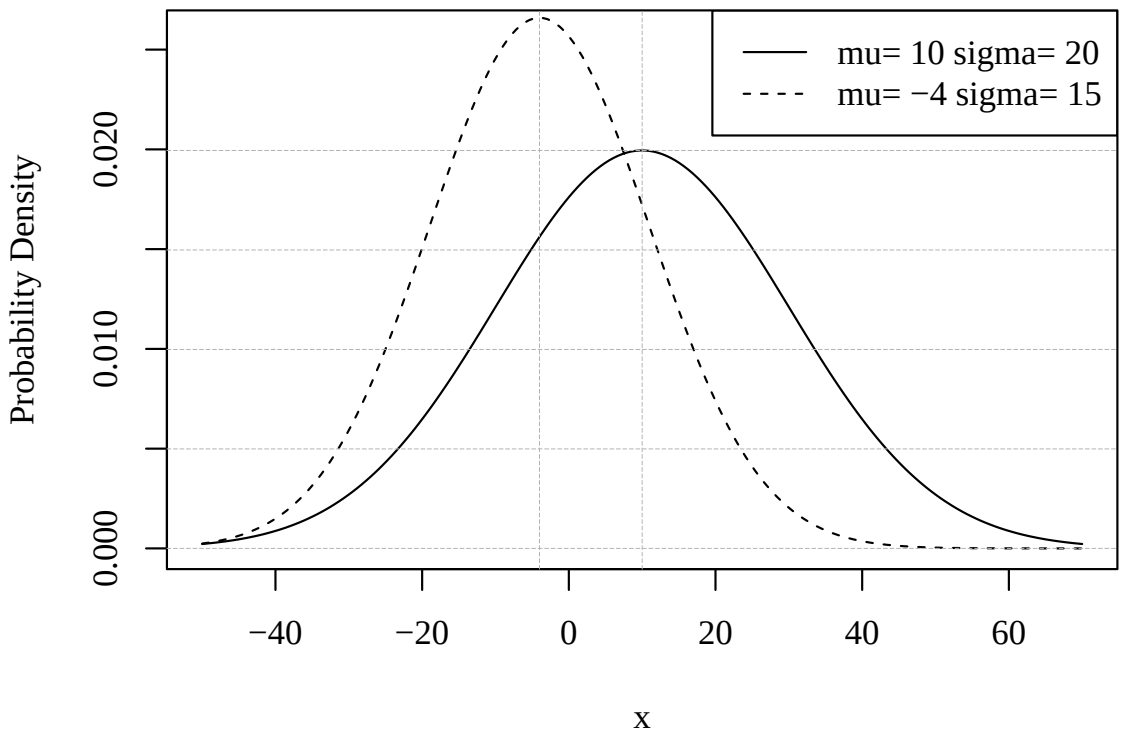
Change μ

- μ shifts normal left and right

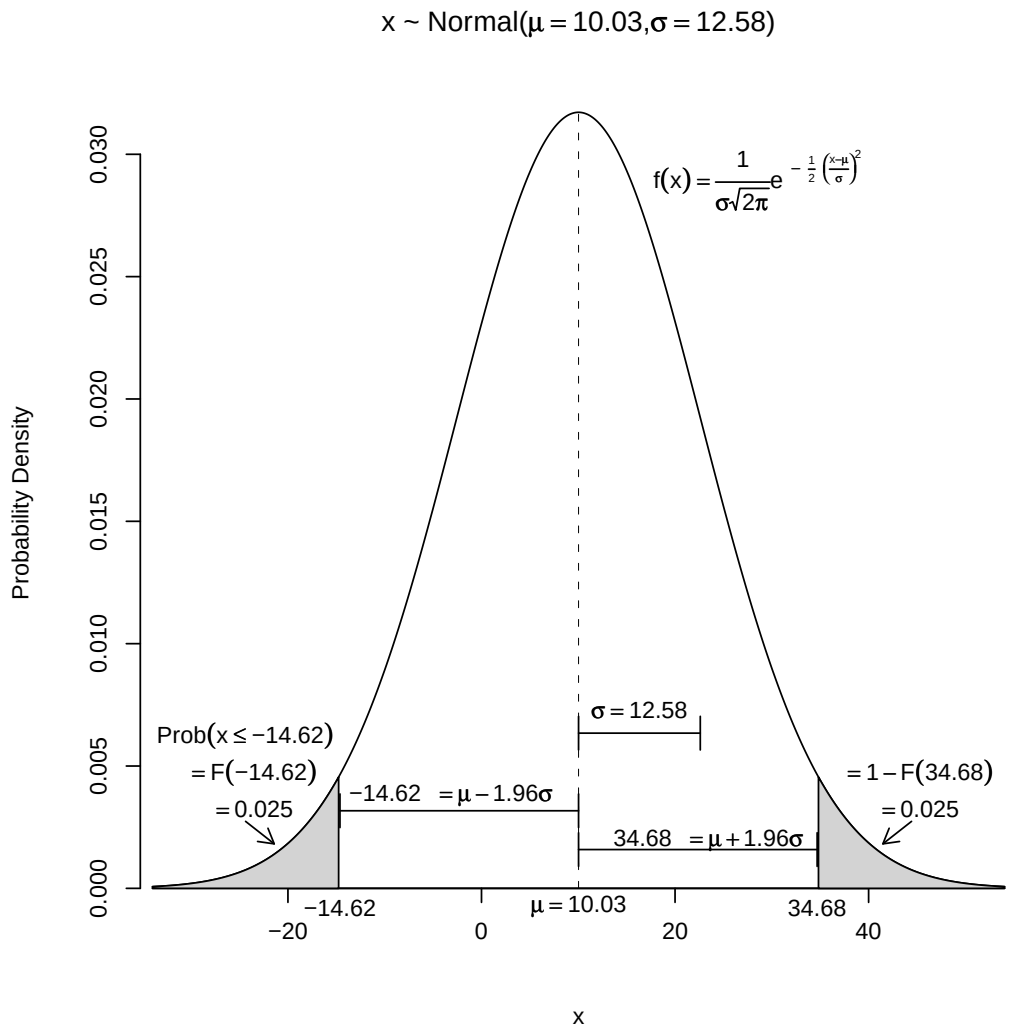


Change μ and/or σ^2

- σ^2 shrinks and stretches it, leaving center "same".
- adjust both



Normal: Most Well-Investigated Distribution



Surprise! Go Looking for Moments, Look what Pops Out.

Suppose $x_i \sim N(\mu, \sigma^2)$. Then

- $E[x_i] = \mu$ (The expected value of x_i is the parameter μ)
- $Var[x_i] = \sigma^2$ (The variance of x_i is the parameter σ^2)

$N(0, 1)$ is called the "Standard" Normal

- Recall Z statistic.

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{29}$$

  $Z_i \sim N(0, 1)$.

- Recover x_i from Z_i

$$x_i = \mu + Z_i \sigma. \tag{30}$$

- Standard Normal Tables in all stat books (used to be...)

CDF not simplify-able (sp?)

Bummer: The CDF does not boil down to some easy formula:

$$F(k;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{k} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{31}$$

Numerical integration required to calculate the chance of outcome above or below a particular point.

Symbol $\hat{\mu}$ (mu hat): an estimate of μ

- Symbol consistency: $\hat{\mu}$ is the estimate of μ.
- The maximum likelihood estimate of μ is the sample average (commonly called $\bar{x}$):

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{32}$$

The maximum likelihood estimate of σ^2 is

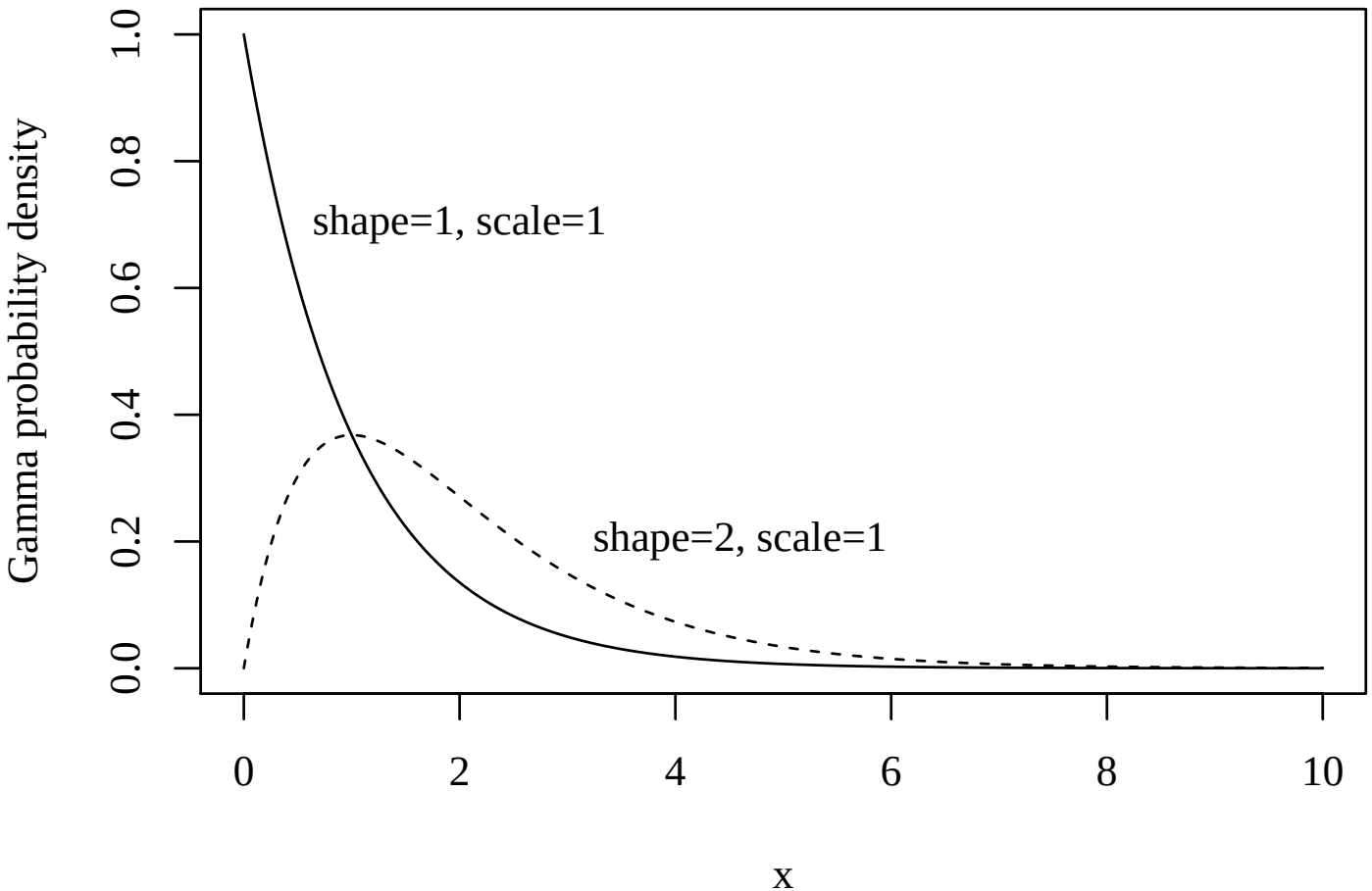
$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})^2}{N}. \tag{33}$$

That ML estimator is biased. But this is not:

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})^2}{N - 1}. \tag{34}$$

Gamma Distribution

May be either "ski-slope" shaped or it may be single-peaked, with a more-or-less exaggerated tail on the right.



Probability Density Function

- $Gamma(\alpha, \beta)$ has parameters: shape (α) and scale (β).
- The PDF is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \text{ where } x \geq 0, \alpha>0, \beta>0. \tag{35}$$

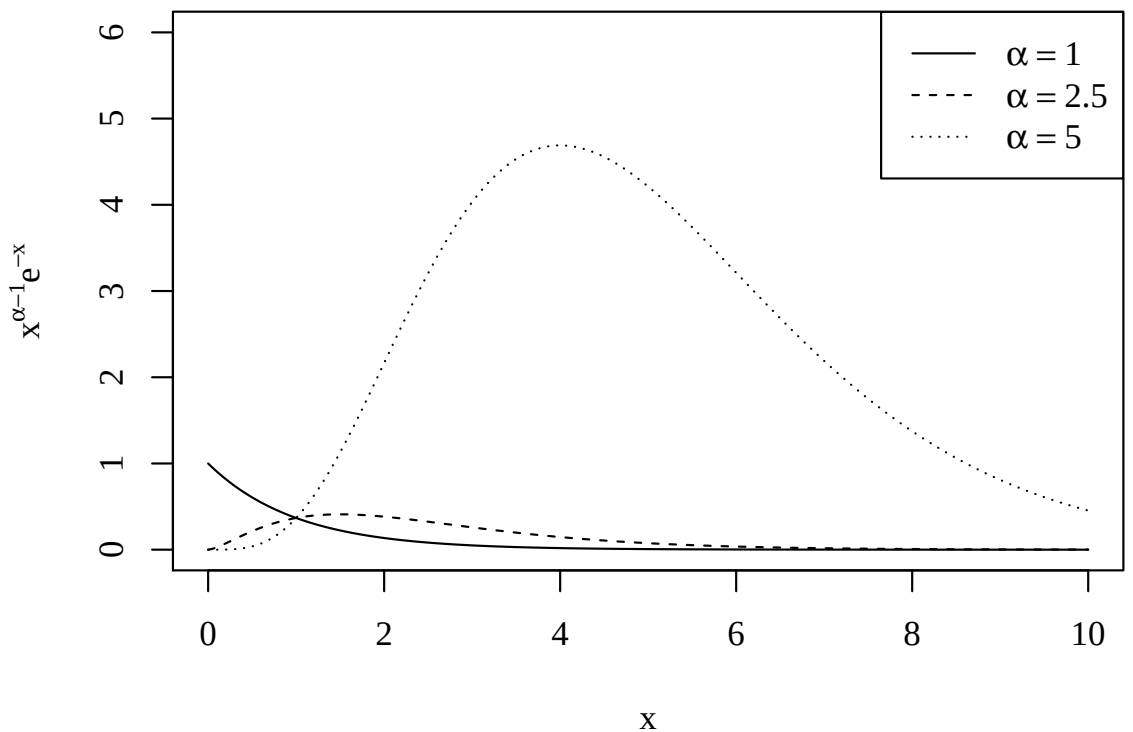
  The symbol $\Gamma(\alpha)$ is a normalizing constant. It is known as the gamma function. It can be thought of as an extension of the factorial function to the real number line. For integers, $\Gamma(\alpha) = (\alpha - 1)!$
- Ignoring the constant part, the kernel of the distribution is

$$x^{\alpha-1} e^{-x/\beta} = \frac{x^{\alpha-1}}{e^{x/\beta}} \tag{36}$$

- So the PDF boils down to "how fast does the numerator grow in comparison to the denominator?"
- The denominator will always win out in the end, causing the density to shrink to 0 as $x \to \infty$.

Why is α the "Shape" Parameter?

- If $\alpha = 1$, this just reproduces the exponential (since $x^0 = 1$).
- If $\alpha > 1$, the shape changes. Its a single-peaked function with a mode in the interior of the domain. That is why α is called a "shape" parameter.



Cumulative Distribution Function

This integral has no "simplified" representation:

$$F(k;\alpha,\beta) = \int_0^k \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx. \tag{37}$$

Numerical approximation required.

Moments

If $x_i \sim Gamma(\alpha, \beta)$,

$$E[x_i] = \alpha \cdot \beta \tag{38}$$

Unlike the Normal, $E[x_i]$ depends on both parameters.

- The variance is more sensitive to the scale parameter.

$$Var[x_i] = \alpha \cdot \beta^2 \tag{39}$$

There's a Mode if...

- If $\alpha > 1$, then the distribution is single-peaked
- And the mode is

$$mode = \beta(\alpha - 1) \tag{40}$$

Links to other distributions

- The $\chi^2(\nu)$ distribution (which is discussed below) has the same PDF as $Gamma(\frac{\nu}{2}, 2)$.
- If $\alpha = 1$, the gamma simplifies into an exponential distribution (24).

Important Properties

- Additivity property. The sum of observations from gamma distributions with various α_i, but the same scale (β), is distributed as $Gamma(\alpha_1 + \ldots + \alpha_n, \beta)$.

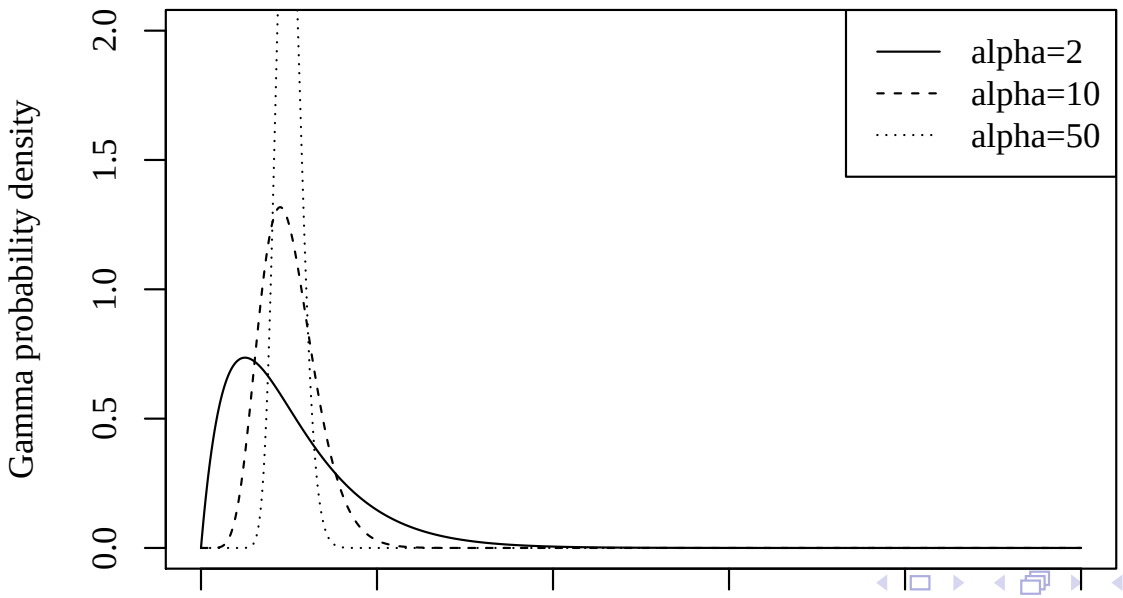
$Gamma(\alpha, \beta)$ is frequently used in "mixture models".

Sometimes we need to add non-negative "noise" that has $E[e_i] = 1$.
Consider $e_i \sim Gamma(\alpha, 1/\alpha)$

$$E[x] = \alpha \cdot \frac{1}{\alpha} = 1. \tag{41}$$

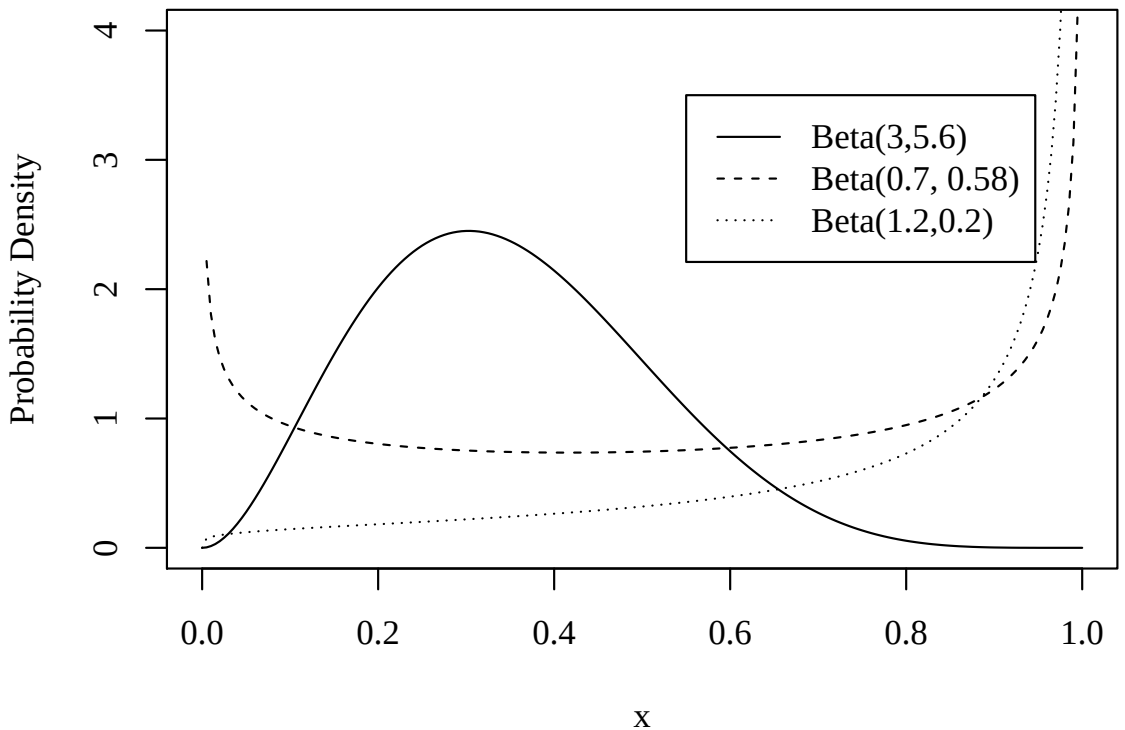
But the variance is flexible.

$$Var[x] = \alpha(\frac{1}{\alpha})^2 = \frac{1}{\alpha}. \tag{42}$$



Beta Distribution

$x_i \sim Beta(\alpha, \beta)$ x_i *is in* [0.1]. Note various PDFs



Probability Density Function

- The standard *Beta*'s pdf is defined on $[0, 1]$:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{43}$$

- and the normalizing constant is called the beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$$

- To get a grasp on this, ignore the constant, focus on $x^{\alpha-1}(1-x)^{\beta-1}$
- Think of α as the "emphasis on 1" and β as the "emphasis on 0".
  - α bigger than β means "big outcomes more likely"
  - β bigger than α means "small outcomes more likely"

Calculate some examples $x^{\alpha-1}(1-x)^{\beta-1}$

$\alpha = 1, \beta = 1$

x	$f(x)$
0.25	1
0.5	1
0.75	1

all equally likely

$\alpha = 0.5, \beta = 1$

x	$f(x)$
0.25	2.0
0.5	1.41
0.75	1.15

prob. declining left to right

$\alpha = 2, \beta = 2$

x	$f(x)$
0.25	0.18
0.5	0.25
0.75	0.187

mode at 0.5

$\alpha = 2, \beta = 1$

x	$f(x)$
0.25	0.25
0.5	0.5
0.75	0.75

big outcomes more likely

$\alpha = 2, \beta = 3$

x	$f(x)$
0.25	0.14
0.5	0.125
0.75	0.047

small outcomes more likely

$\alpha = 4, \beta = 2$

x	$f(x)$
0.25	0.012
0.5	0.065
0.75	0.105

Inter-Linkages

- fraction formed by two gamma variables that have the same scale parameter, $x_1/(x_1 + x_2)$, is distributed as a beta variable.

Cumulative Distribution Function

The chance that a draw from a beta density is less than k is

$$F(k;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \int_0^k x^{\alpha-1}(1-x)^{\beta-1} dx \tag{44}$$

Moments

If $x_i \sim Beta(\alpha, \beta)$,then:

$$E[x] = \mu = \frac{\alpha}{\alpha + \beta} \tag{45}$$

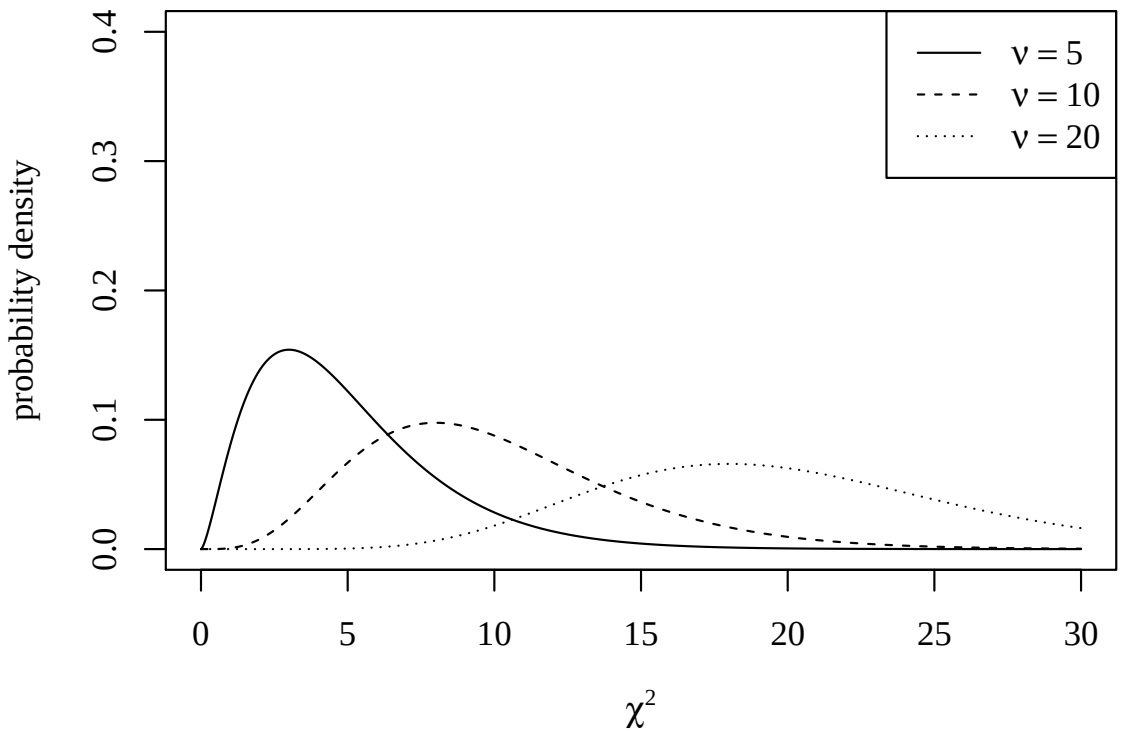
$$Var[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{46}$$

If $\alpha > 1$and $\beta > 1$, the mode of the *Beta* distribution is

$$mode = \gamma = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{47}$$

Chi-Squared

The Chi-Squared may be referred to as $\chi^2(\nu)$ or χ^2_ν .
The Chi-Squared distribution is used to describe the "sum of squared mistakes" or "mismatches" between expectation and observation.



Probability Density Function

Recall, the pdf of a Chi-square distribution is identical to a gamma distribution with shape parameter $\nu/2$ and scale 2 (35). Thus Chi-square's probability density function is

$$f(x) = \frac{1}{\Gamma(\frac{\nu}{2})(2)^{\frac{\nu}{2}}} x^{\frac{\nu}{2}-1} e^{-x/2}, \; x \geq 0, \; \nu > 0. \tag{48}$$

Surprise: sum of squared random variables follows a Chi-square distribution.

Draw a collection of ν observations from a standard normal distribution,

$$Z_i \sim N(0,1), \text{ for } i = 1, 2, \ldots, \nu. \tag{49}$$

Square each one, and add them together. The result is distributed as a $\chi^2(\nu)$. That is to say

$$Z_1^2 + \ldots + Z_\nu^2 \sim \chi^2(v). \tag{50}$$

When it is used in this context, the parameter that represents sample size, ν, is often called "degrees of freedom."

Moments

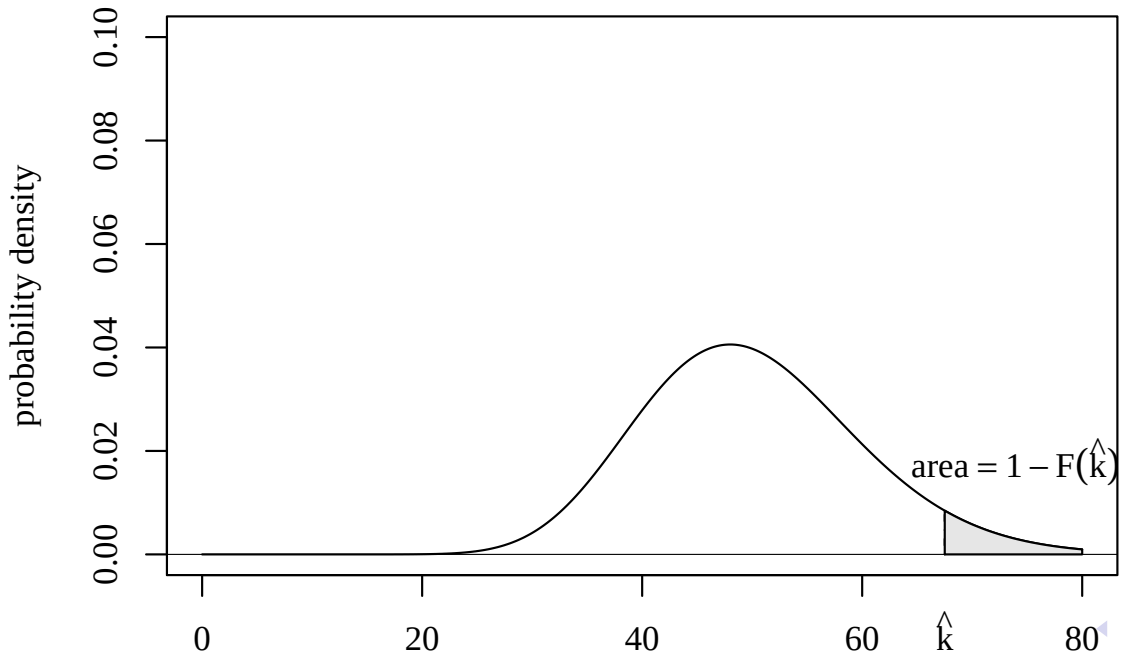
Since the $\chi^2(\nu)$ is the same as $Gamma(\frac{\nu}{2}, 2)$,

$$E[x] = \nu \tag{51}$$

$$Var[x] = 2\nu \tag{52}$$

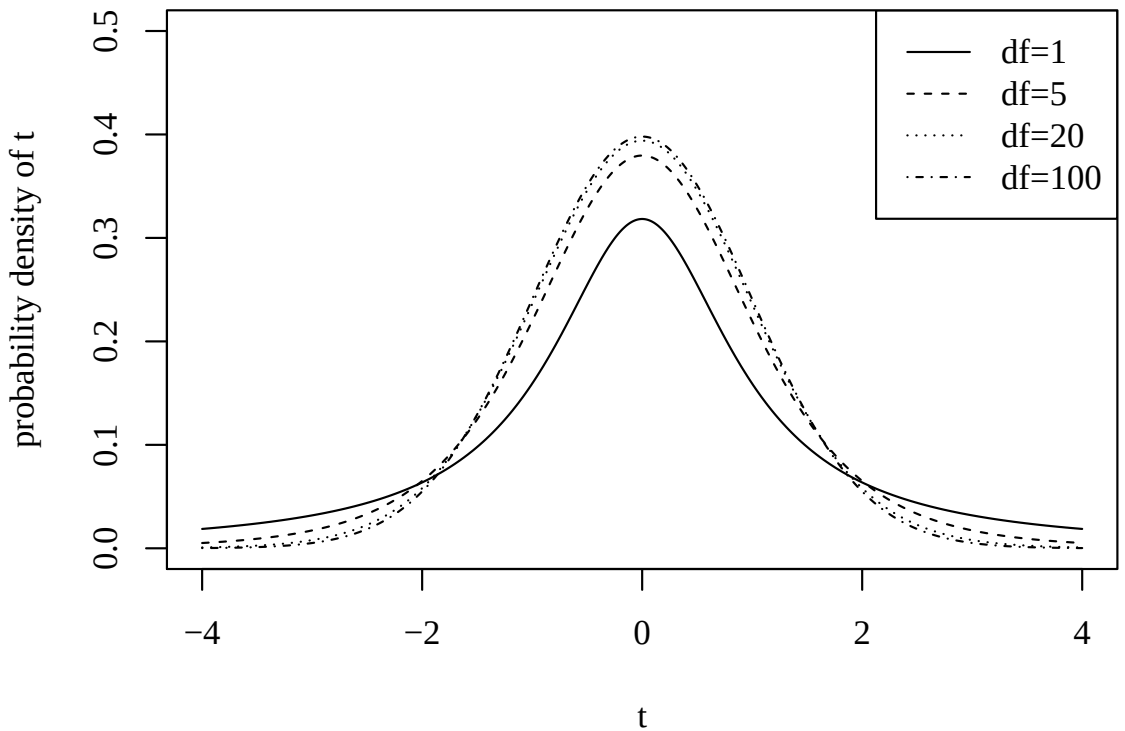
Chi-Square in Model Comparison Test

- Many statistical procedures can result in a estimate that is distributed as $\chi^2(\nu)$.
  - The mis-match between the saturated model and the fitted generalized linear model, for example, is distributed as a χ^2.
- The top 5% under the pdf of $\chi^2(50)$ is drawn. The shaded area on the right–values greater than 67.50–represents the top 5% of possible draws from $\chi^2(50)$.



Student's t distribution

- symmetric and uni-modal.
- one parameter, ν, known as "degrees of freedom" (df).
- Similar to Normal(0,1)
- Statisticians say that t has "fatter tails" the normal.



Probability Density Function

The probability density of the t distribution is

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\frac{\nu+1}{2})}. \tag{53}$$

- Sorry, I got no intuition from that!
- Its center point–expected value, median, and mode–is 0.

t is a Workhorse in hypothesis testing

- The end result will concern a parameter θ.

$$\frac{\hat{\theta} - E[\theta]}{standard\ error(\hat{\theta})} \sim t(\nu).$$

- Z stat intuition: If we only knew the standard deviation of the mean, we could calculate a thing-like-a-Z-statistic.

$$\frac{estimated\ mean - null\ hypothesis}{standard\ deviation\ of\ mean}.$$

- A standard normal variable could be created if we knew the true variance, as in

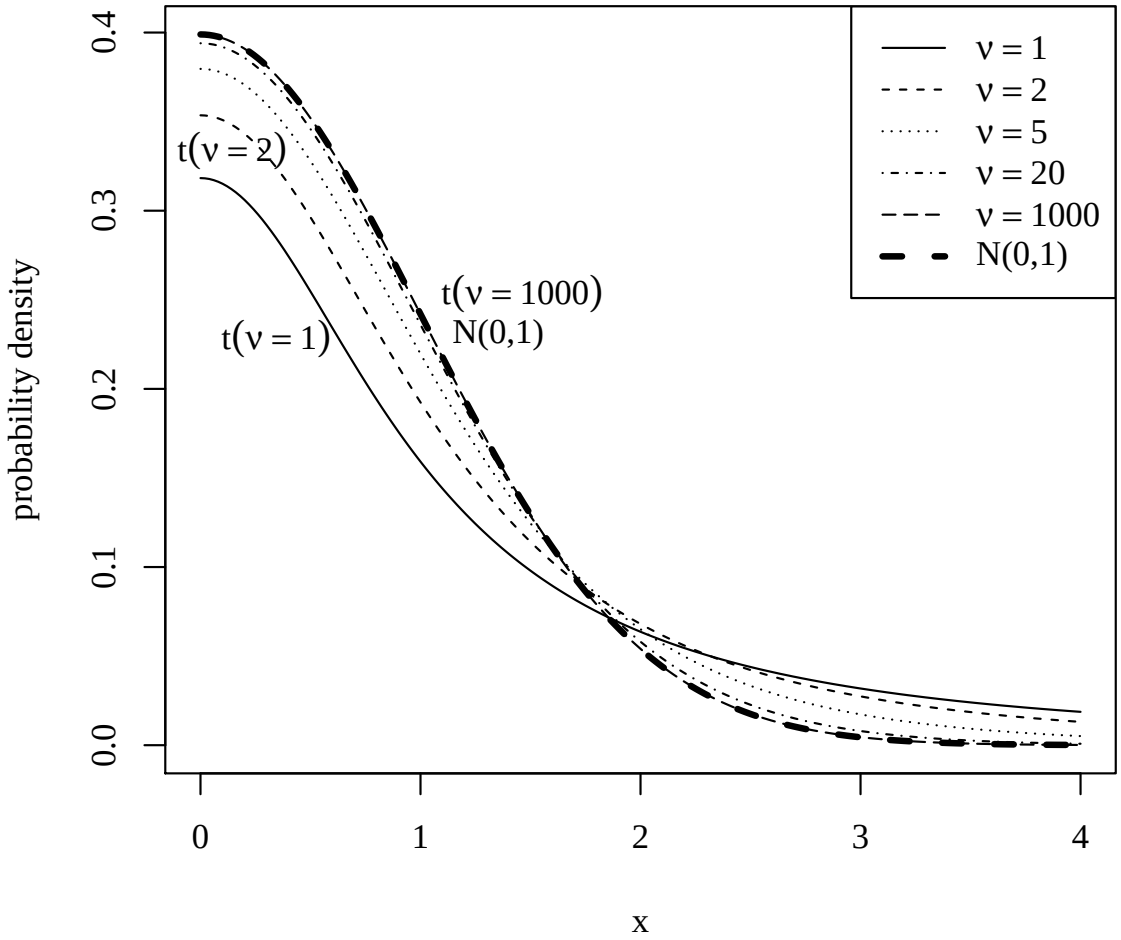
$$\frac{\widehat{E[x]} - E[x]}{\sqrt{Var[x]/N}} \sim N(0,1)$$

- t is a work-around for problem that $Var[x]$ is unknown.
- Replace $Var[x]$ by estimate $\widehat{Var[x]}$, and so $\sqrt{\widehat{Var[x]}} = \widehat{StdDev}[x]$

$$"t\ ratio" = \frac{\widehat{E[x]} - E[x]}{\sqrt{N}\widehat{StdDev}[x]}.$$

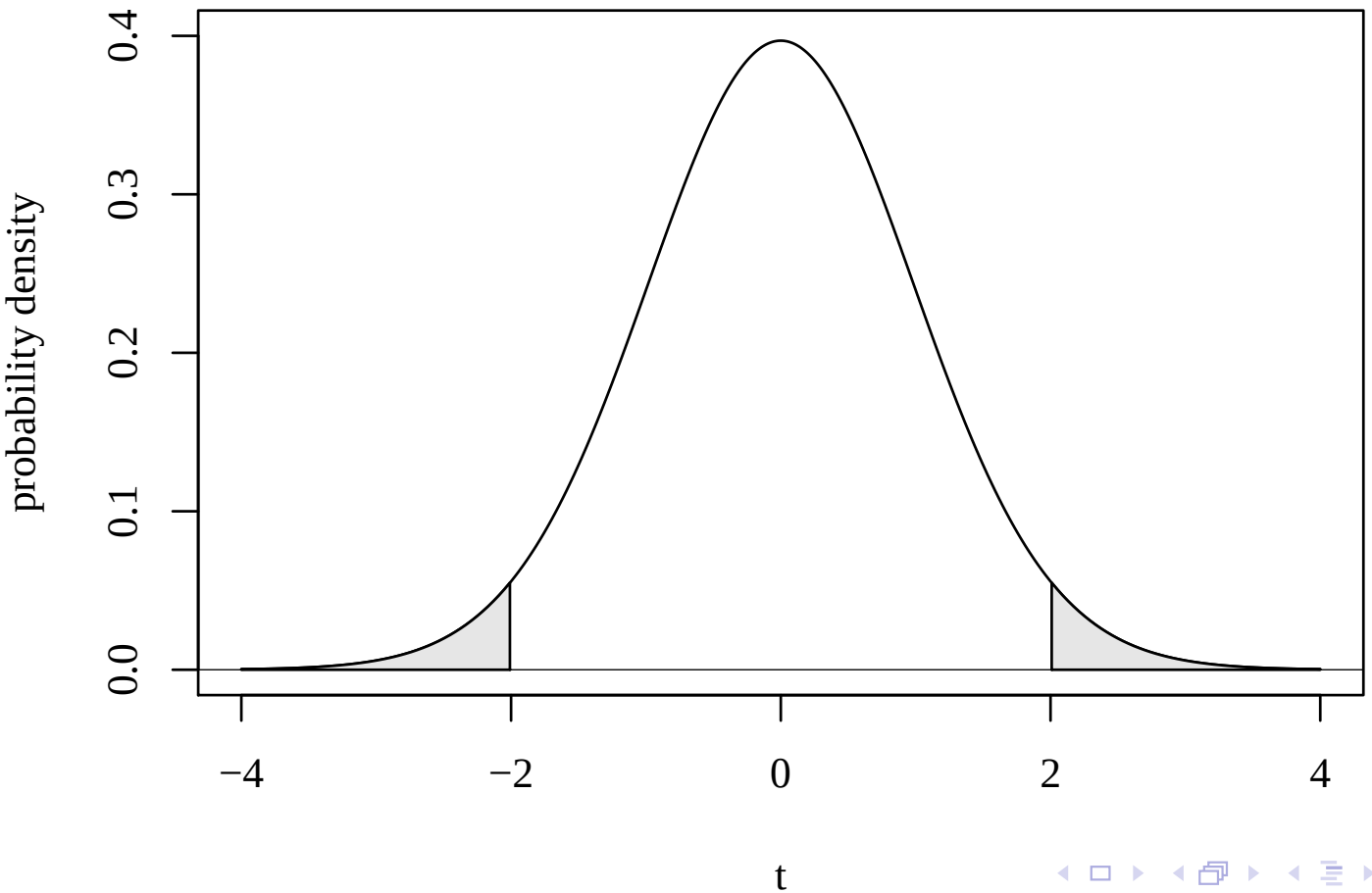
$t \to$ Normal(0,1) as $\nu \to \infty$

When a sample is large, then the t ratio and the standard normal (29) are not noticeably different.



t is Often Interpreted as a Two Tailed Distribution

Unlike the χ^2 distribution, where we look only on the right tail of the distribution for evidence of unusual cases, the t distribution has critical regions both tails.



Moments

- Supposing $\nu \geq 1$, the expected value, median, and mode of a t distribution are all 0.
- The variance of a t distribution is

$$Var[x] = \frac{\nu}{\nu - 2} \tag{54}$$

- Note: as $\nu \to \infty$, $Var[x] \to 1.0$, consistent with the claim that the t density converges to $N(0, 1)$.

Comments

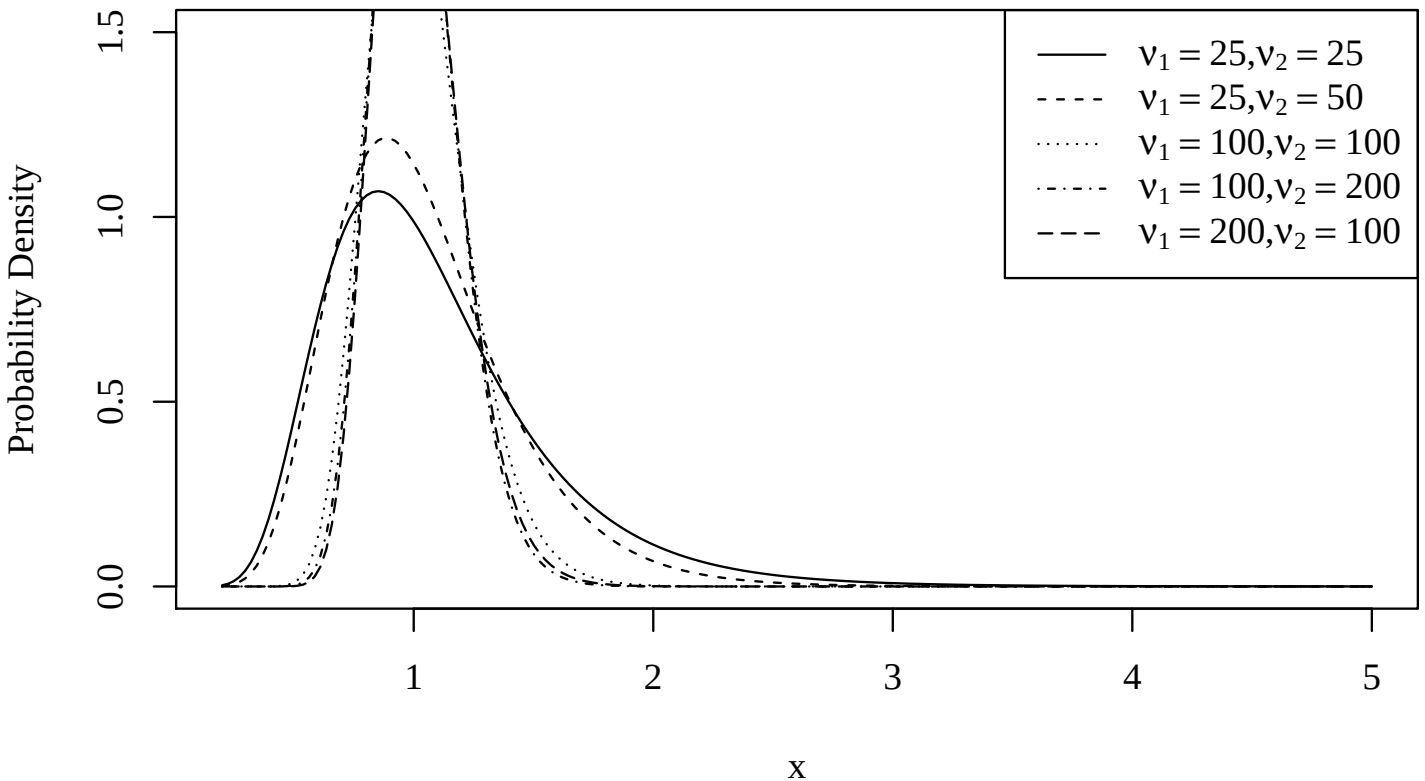
The t distribution a work horse in everyday statistics.
Many estimators boil down to a comparison of an estimate and its standard error.
"t ratio" commonly refers to any Normally distributed estimator, $\hat{\theta}$, against its standard error.

$$\frac{\hat{\theta}}{std.error(\hat{\theta})} \tag{55}$$

(More on hypothesis testing later...)

The F distribution

The $F(\nu_1, \nu_2)$ distribution ("F" is for Fisher) describes a variable on $[0, \infty)$. It depends on 2 parameters, ν_1 and ν_2.



PDF is difficult to comprehend.

$$f(x;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}x^{\frac{\nu 1}{2}-1}\left(\nu_2+\nu_1 x\right)^{-(\nu_1+\nu_2)/2} \tag{56}$$

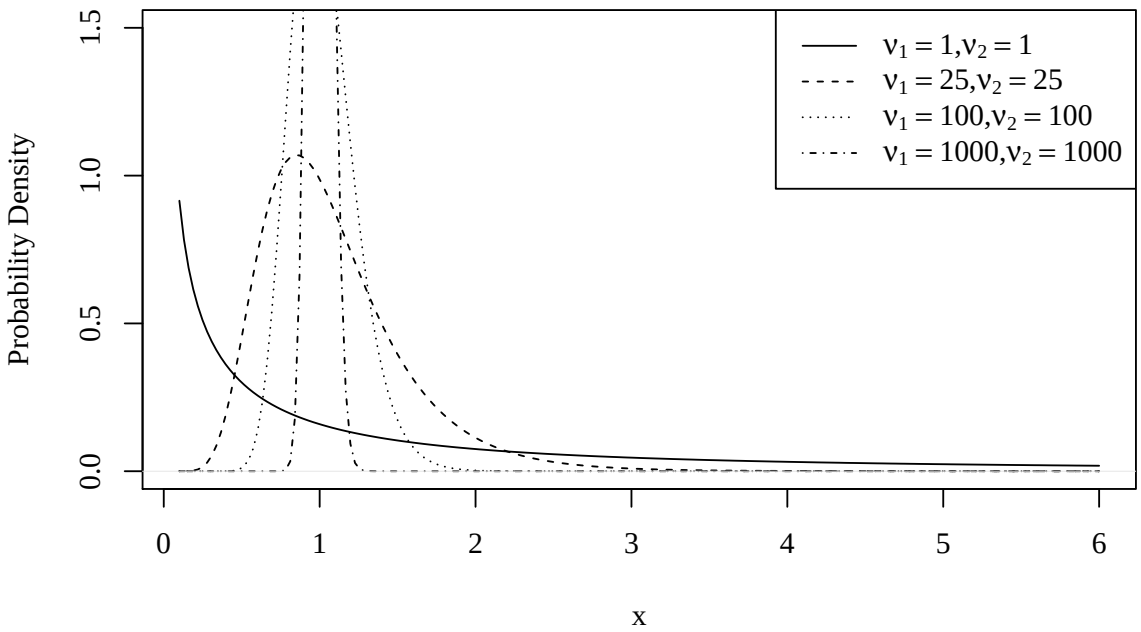
"Tell Me A Story" Instead

- $Z_1^2 + Z_2^2 + Z_{\nu_1}^2$ is distributed as a $\chi^2(\nu_1)$.
- Compare against a second set of observations, $Z_1^2 + Z_2^2 + Z_{\nu_2}^2 \sim \chi^2(\nu_2)$.
- So far as I know, there is no method to compare the difference of two χ^2 statistics, but it is possible to compare their ratio.
- The test statistic we want to understand is thus a ratio of "mean squares":

$$\frac{\textit{Sample}\,1: \quad (Z_1^2 + Z_2^2 + \ldots + Z_{\nu_1}^2)/\nu_1}{\textit{Sample}\,2: \quad (Z_1^2 + Z_2^2 + \ldots + Z_{\nu_2}^2)/\nu_2} \textit{ is distributed as } F(\nu_1, \nu_2). \tag{57}$$

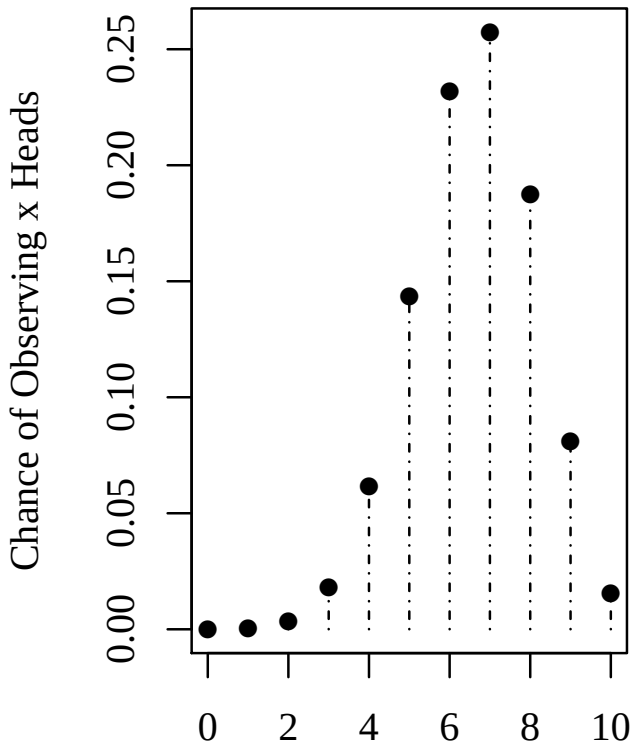
  The pdf of $F(\nu_1, \nu_2)$ represents the diversity we would observe if we repeatedly drew ν_1 and ν_2 observations and then formed this ratio of mean squares.
- If $\nu_1 = 1$, then the density of F is the same as that of squared t variable.

Intriguing: pdf of F Collapses Around 1 as Sample Size Increases

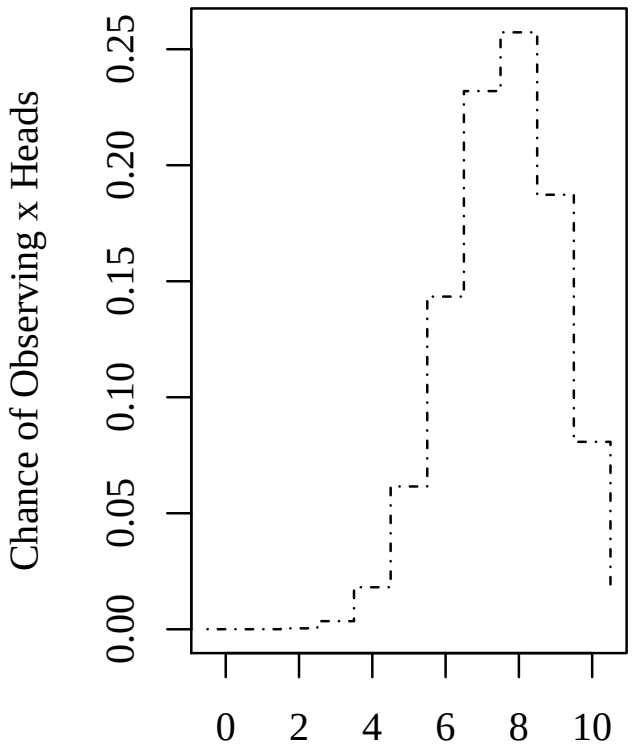


Binomial Distribution

$B(N, \pi)$ the number of "events" (or "successes", or "wins", etc.) when there are N "trials" and the chance of a success on each trial is fixed at π.



10 Flips with a Biased Coin



10 Flips with a Biased Coin

Probability Mass Function

$$Prob(k|N,\pi) = \frac{N!}{(N-k)!k!}\pi^k(1-\pi)^{N-k} \tag{58}$$

1. If there are N independent trials, and how likely we are to get k successes. The chance that the first k trials will succeed, and the rest will fail, is

$$\pi \times \pi \times \{k \text{ times}\} \times (1-\pi) \times (1-\pi) \times \{N-k \text{ times}\} = \pi^k(1-\pi)^{N-k}$$

2. $\frac{N!}{(N-k)!k!}$ is the number of ways to re-arrange N things so that k are successes and $N-k$ are not.

Example

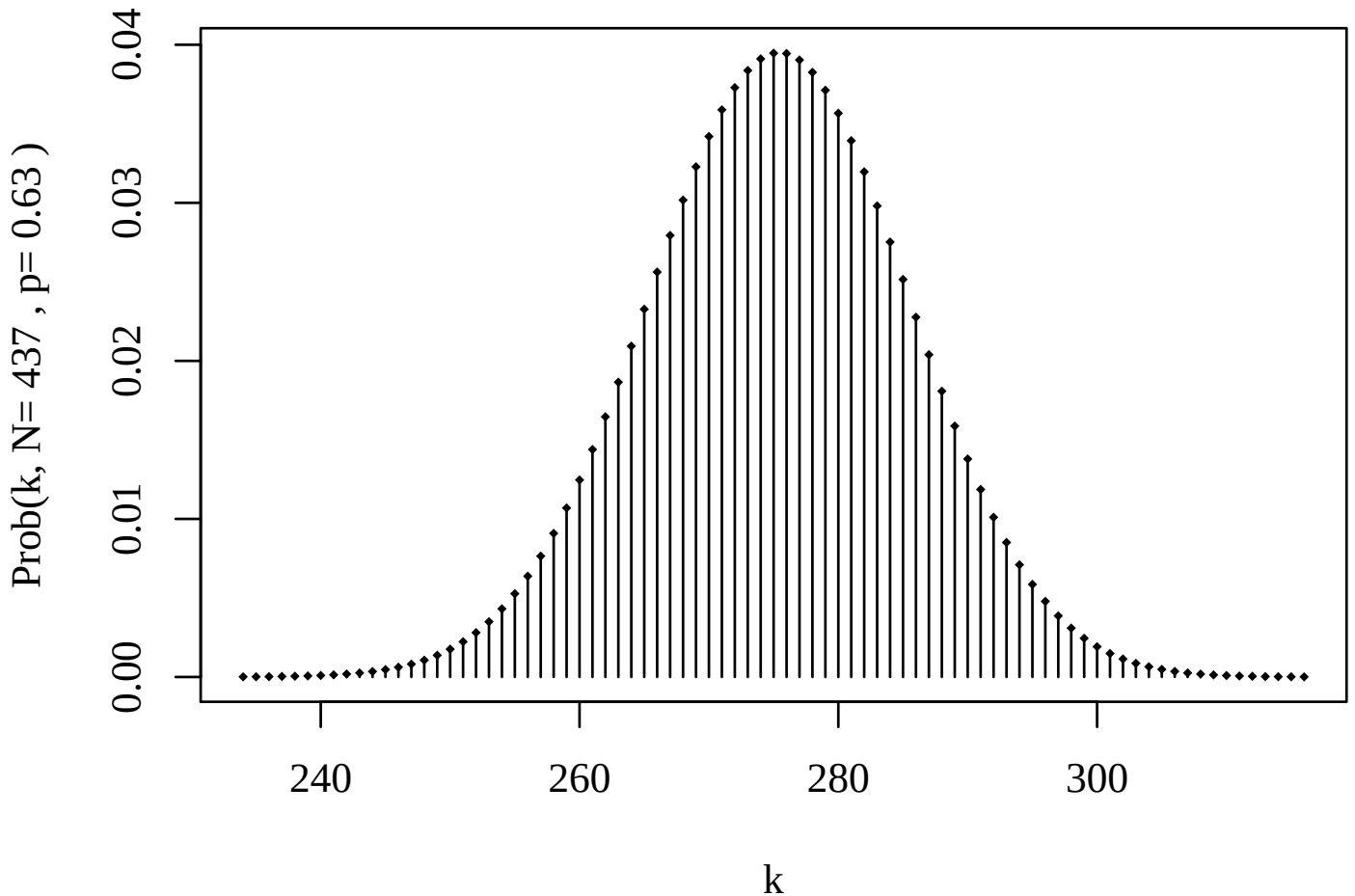
- Consider 437 women having babies, the chance of having a boy baby is 0.63.
- The chance of k boys is:

$$Prob(k|437, 0.63) = \frac{437!}{(437-k)!k!}(0.63)^k(1-0.63)^{437-k} \tag{59}$$

- The probability of 300 boys:

$$Prob(300|437, 0.63) = 0.000112 \tag{60}$$

Binomial is Intriguingly Normal if N is large



Moments

The expected value is:

$$E[x] = \pi \cdot N \tag{61}$$

and the variance is

$$Var[x] = \pi(1 - \pi)N \tag{62}$$

Derivation: Based on "Sum of Bernoulli Trials" Interpretation

For instance, an observed sample is N "Bernoulli trials", $\{x_1, x_2, \ldots, x_N\}$, such as

$$0, 1, 1, 0, 1, 1, 0, 0 \ldots, 1, 0 \tag{63}$$

The number of successes is the sum of those trials

$$x_1 + x_2 + x_3 + \ldots + x_{N-1} + x_N \tag{64}$$

Each x_i is a "Bernoulli trial, and obviously

$$E[x_1] = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi \tag{65}$$

So the Binomial expected value

$$\begin{aligned} E[x_1 + x_2 + \ldots x_N] &= E[x_1] + E[x_2] + \ldots + E[x_n] \\ &= \pi + \pi + \ldots + \pi \\ &= N \cdot \pi \end{aligned} \tag{66}$$

Variance

Consider one Bernoulli trial, x_1, in isolation. Its variance is

$$\begin{aligned} Var[x_1] &= \pi(1 - E[x_1])^2 + (1 - \pi)(0 - E[x_1])^2 \\ &= \pi(1 - \pi)^2 - (1 - \pi)(-\pi)^2 \\ &= \pi(1 - 2\pi + \pi^2) + \pi^2 - \pi^3 \\ &= \pi - 2\pi^2 + \pi^3 + \pi^2 - \pi^3 \\ &= \pi - \pi^2 = \pi(1 - \pi) \end{aligned} \tag{67}$$

Treat Binomial as sum of N independent trials (so Covariance=0). Thus, the law for calculating the variance of a sum of terms applies.

$$\begin{aligned} Var[x_1 + x_2 + \ldots x_N] &= Var[x_1] + Var[x_2] + \ldots + Var[x_N] \\ &= \pi(1 - \pi) + \pi(1 - \pi) + \ldots + \pi(1 - \pi) \\ &= \pi(1 - \pi)N \end{aligned} \tag{68}$$

How Does This Arise in Regression?

- Consider groups of test subjects.
- Dependent variable is number of 'successes' out of N_j respondents in group j.
- Suppose $Binomial(N_j, \pi_j)$. N_j is known (because of design), we need to predict π_j as a function of parameters and independent variables.

Poisson Distribution: Event Count Model

The Poisson is a discrete distribution, most commonly for "event counts" on $0, 1, \ldots, \infty$.
Poisson represents a process characterized as:

- The chance of one event during the passage of time Δt is approximately $\lambda \cdot \Delta t$ (and, as Δt shrinks to 0, $\lambda \Delta t$ approximates the chance of an event more and more closely).
- The chance of a second event in a particular chunk of time is vanishingly small.

PMF

- One parameter, λ.

$$f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!},\ where\ x \geq 0, \lambda > 0 \tag{69}$$

- The term $e^{-\lambda}$ (same as $1/e^{\lambda}$) is a normalizing constant.
- The kernel of this probability model is simply

$$\frac{\lambda^x}{x!} \tag{70}$$

Any Sequence Can Be the Backbone of a PMF

- Recall that "any integrable function" can be backbone for a PDF?
- For discrete models, "any convergent sequence" can be the backbone of a PMF. That is, if $S = \sum p(x_i)$ exists, then the PMF can be $\frac{1}{S} p(x_i)$
- Poisson example

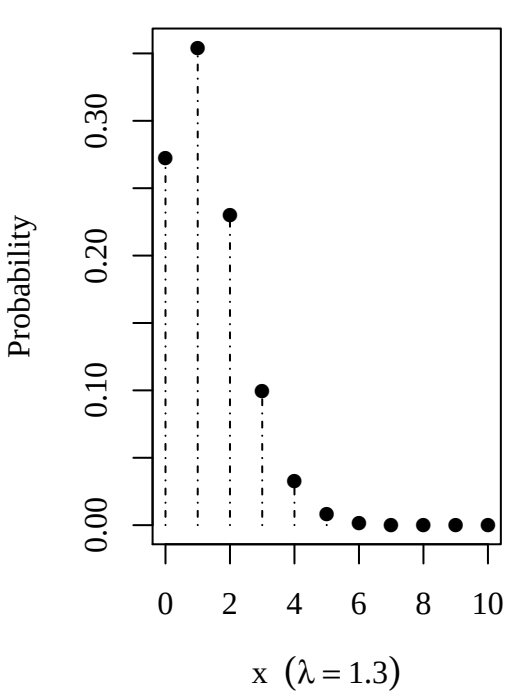
x		0	1	2	3	4	5	...	∞
$\lambda^x/x!$		1	λ^1	$\lambda^2/2!$	$\lambda^3/3!$	$\lambda^4/4!$	$\lambda^5/5!$		$\lambda^\infty/\infty!$

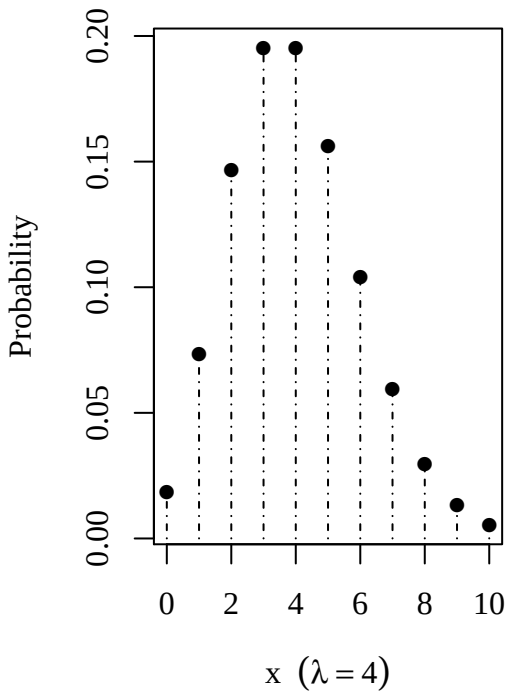
The sum of the items in the second row is

$$exp(\lambda) = 1 + \lambda + \lambda^2/2! + \lambda^3/3! + \lambda^4/4! + \lambda^5/5! + \ldots + \lambda^\infty/\infty! \quad (71)$$

Shape of the PMF

- If $\lambda < 1$, then the most likely outcome is always 0 and higher values are progressively less likely.
- If $\lambda > 1$, the mode shifts into the interior.





Moments

The expected value is equal to its variance, and both of them are equal to λ.

$$E(x) = \lambda$$

$$Var(x) = \lambda$$

How Does this Arise in Regression?

- Think about a set of observed counts, $y_1, y_2, \ldots, y_N$.
- Build a model that treats each as a draw from its own "customized" Poisson distribution.

$$f(y_i; \lambda_i) = e^{-\lambda_i} \frac{\lambda^{y_i}}{y_i!} \tag{72}$$

- Build a "predictive model" that depends on parameters β_j and independent variables:
- $$\lambda_i = exp(\beta_0 + \beta_1 z_i) \tag{73}$$

  Try to estimate the parameters β_0 and β_1

Outline

1 What is Probability?
2 Characterizing Distributions
- Expected Value
- Variance

3 Algebra of Expected Values and Variances
4 Example Distributions
- Exponential
- Normal
- Gamma
- Beta
- χ^2 (Chi-Squared)
- t
- F
- Binomial
- Poisson

5 Practice Problems

Problems

1. The expected value of the height of a black bear is 5 feet. The expected value of the width of a black bear is 6 feet. You collected lots of data on the height and widths of bears during a long, boring field study. You are excited to know if the bears you study are like other bears that people have already studied. Before leaving the camp, you sent the data home to your assistant. Your research assistant mistakenly added the two columns of number together. You are in the "publish or perish" world, and you have to make something out of that data. Your numbers are the sum of height and width, you decide to call your new random variable the "index of bear displacement (IBD)."
    1. What is the expected value of the IBD, supposing your bears are like other bears people have studied before? How do you figure that out when I don't give you a column of numbers?
    2. If I gave you a column of numbers for your bears, would it help you to calculate the expected value?

Problems ...

3. On a conceptual level, how would you describe the difference between the expected value of the IBD and the sample average of your accidentally-added-together column of data?

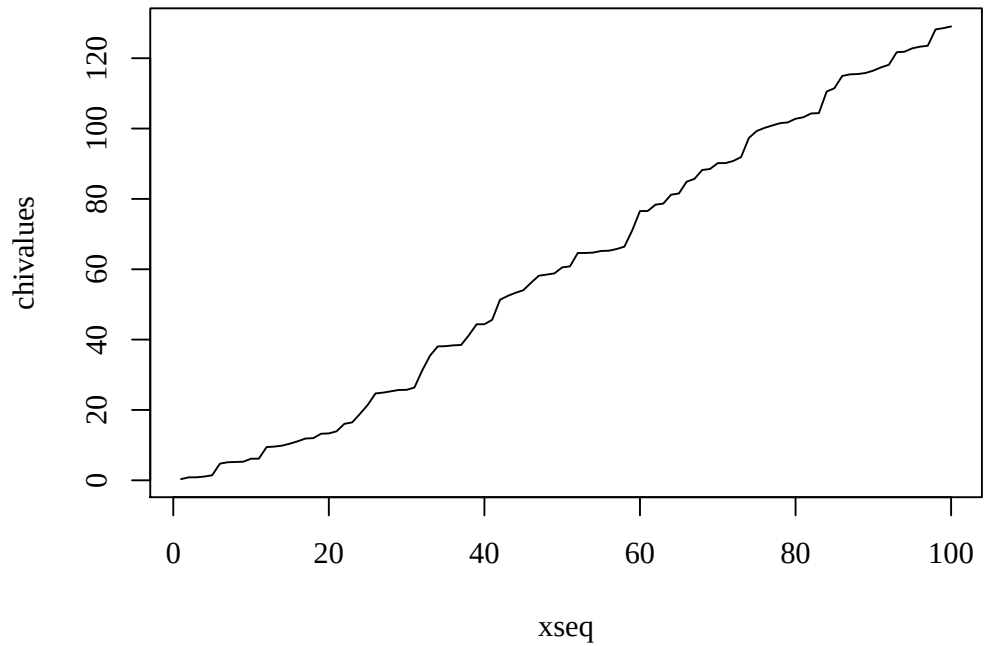
2. Note one can convert from Celsius to Fahrenheit with $F = 32 + \frac{9}{5}C$
   1. Your professor asks you to calculate the expected value of the temperature today. You get out some complicated probability model and work out the answer. The Expected Value of the temperature 18 Celsius. But the professor wanted the temperature in Fahrenheit, and you are too busy to go back and re-do all of the pdf calculations to get that for her. There's a shortcut to convert your result of 18C to Fahrenheit. How do you use it?
   2. This might force you to use your head. If the variance in Celsius is 4, what would be the variance of the same data presented in Fahrenheit? You have to use your head, but I'll give you a hint. The variance of a constant is always 0. So $Var[32] = 0$. Using that fact, I bet you can use the result stated in **??**.

Problems ...

3. I collected 1300 observations from a random process. I'm not sure which distribution was used, I forgot to write it down. But my teacher told me that E[x]=3 and Var[x]=10. The teacher told me to calculate the sample average and find out how certain she should be about that estimate. She said calculate Variance of the average. But I can't understand how to calculate the variance when I don't know to calculate $Var[\bar{x}]$. So I am completely stumped. Can you help?
4. The sum of squared observations from a Normal distribution takes on the form of a Chi-Square distribution. Lets use R to check that out. Here's some code that use R to create 100 normal observations, and then squares them, and then calculates the cumulative sum. Then it makes a plot.

```
xseq <- 1:100
x <- rnorm(100)
xsquare <- x*x
chivalues <- cumsum(xsquare)
plot(xseq, chivalues, type="l")
```

Problems ...



1. Think for a while about what that plotted line means. Then write a paragraph.
2. Run that program over 10 or 20 times, compare the plots you get. What do the differences among them signify?

Problems ...

5. Lets beat up a Gamma variable for a while. Suppose I give you a Gamma distributed variable with shape $= \alpha$ and scale $= \beta$.
    1. What is the Expected Value and the Variance?
    2. If α is replaced by a new shape parameter $2/\beta$, calculate E and Var
    3. Use R to sketch the pdf of Gamma(x; shape=2, scale=3)
    4. Use R to tell me the probability density of a particular value of x, say 0.5.
    5. Use R to tell me the chance of observing a value between 0.5 and 1.4