Probability and Sampling

EH6127 – Quantitative Methods

Steven V. Miller

Department of Economic History and International Relations

Stockholm University

Goal(s) for Today

1. Introduce some basic features of probability, counting, and the normal distribution.
2. Simulate central limit theorem and the logic of random sampling.

Probability

Probability refers to the chance of some event occurring.

- It's a ubiquitous feature of the world and you should know it anyway.
- Interestingly, it was developed rather late in human history.
- Origins: gambling in the 17th-18th centuries.

Probability theory is a precursor to statistics and applied mathematics.

- It's mathematical modeling of uncertain reality.

Rules of Probability

Here are some (but not all) important rules for probability.

1. Collection of all possible events (E_1 ... E_n) is a **sample space**.
   - S as a **set** for a coin flip S = { Heads, Tails }.
2. Probabilities must satisfy inequality $0 \leq p \leq 1$.
3. Sum of probability in sample space must equal 1.
   - Formally: $\Sigma_{E_i \in S} \; p(E_i) = 1$
4. If event A and event B are *independent* of each other, the **joint probability** of both occurring is $p(A, B) = p(A) * p(B)$.
5. If probability of event A depends on event B having already occurred, the **conditional probability** of A "given" B is a bit different.

$$p(A \mid B) = \frac{p(A, B)}{p(B)}$$

Total Probability and Bayes' Theorem

Recall: $p(A \mid B) = \frac{p(A,B)}{p(B)}$. And: $p(B \mid A) = \frac{p(B,A)}{p(A)}$

- Moving stuff around: $p(A, B) = p(A \mid B) * p(B)$ and $p(B, A) = p(B \mid A) * p(A)$.
- By definition: $p(A, B) = p(B, A)$
- Therefore: $p(A \mid B) * p(B) = p(B \mid A) * p(A)$

Two different total probabilities emerge:

- $p(A \mid B) = \frac{p(B \mid A) * p(A)}{p(B)}$
- $p(B \mid A) = \frac{p(A \mid B) * p(B)}{p(A)}$

Confusing the two as equivalent is what you call a **prosecutor's fallacy**.

- Read about Sally Clark for a real life horror story of the misuse of conditional probability.

Counting

A basic premise to computing probability is counting.

- It seems basic, but there are multiple ways of doing this.

There's a thing called the **Fundamental Theorem of Counting**:

1. If there are *k* distinct decision stages to a process...
2. ...and each has its own n_k number of alternatives...
3. ...then there are $\prod_{i=1}^{k} n_k$ possible outcomes.

Four Methods of Counting

A form of counting follows choice rules of **ordering** and **replacement**.

1. Ordered, with replacement (n^k alternatives)
2. Ordered, without replacement ($\frac{n!}{(n-k)!}$ alternatives)
3. Unordered, without replacement ($\frac{n!}{(n-k)!k!} = \binom{n}{k}$ alternatives)

There's a fourth method (unordered, with replacement), but it is unintuitive, not much used, and I won't belabor it here.

An Illustration

Assume a class of 25 students. How many different ways can I devise a random sample of 5?

- Ordered, with replacement: 25^5 = 9,765,625
- Ordered, without replacement: $\frac{n!}{(n-k)!} = \frac{25!}{20!}$ = 6,375,600
- Unordered, without replacement: $\binom{25}{5}$ = 53,130

Binomial Theorem

The most common use of a choose notation is the **binomial theorem**.

- Given any real numbers *X* and *Y* and a nonnegative integer *n*,

$$(X+Y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{1}$$

A special case occurs when $X = 1$ and $Y = 1$.

$$2^n = \sum_{k=0}^{n} \binom{n}{k} \tag{2}$$

Binomial Theorem

The binomial expansion increases in polynomial terms at an interesting rate.

$$\begin{aligned}
(X+Y)^0 &= 1 \\
(X+Y)^1 &= X+Y \\
(X+Y)^2 &= X^2+2XY+Y^2 \\
(X+Y)^3 &= X^3+3X^2Y+3XY^2+Y^3 \\
(X+Y)^4 &= X^4+4X^3Y+6X^2Y^2+4XY^3+Y^4 \\
(X+Y)^5 &= X^5+5X^4Y+10X^3Y^2+10X^2Y^3+5XY^4+Y^5
\end{aligned} \tag{3}$$

Notice the symmetry?

Pascal's Triangle

The coefficients form **Pascal's triangle**, which summarizes the coefficients in a binomial expansion.

$n = 0$:						1					
$n = 1$:					1		1				
$n = 2$:				1		2		1			
$n = 3$:			1		3		3		1		
$n = 4$:		1		4		6		4		1	
$n = 5$:	1		5		10		10		5		1

These Have a Purpose for Statistics

Let's start basic: how many times could we get heads in 10 coin flips?

- The sample space $S = \{ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \}$
- We expect 10 heads (or no heads) to be unlikely, assuming the coin is fair.

Binomial Mass Function

This is a combination issue.

- For no heads, *every* flip must be a tail.
- For just one head, we have more combinations.

How many ways can a series of coin flips land on just one head?

- For a small number of trials, look at Pascal's triangle.
- For 5 trials, there is 1 way to obtain 0 heads, 5 ways to obtain 1 head, 10 ways to obtain 2 and 3 heads, 5 ways to obtain 4 heads, and 1 way to obtain 5 heads.

Binomial Mass Function

This is also answerable by reference to the **binomial mass function**, itself derivative of the **binomial theorem**.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{4}$$

where:

- x = the count of "successes" (e.g. number of heads in a sequence of coin flips)
- n = the number of trials.
- p = probability of success in any given trial.

Binomial Mass Function

What's the probability of getting five heads on ten fair coin flips.

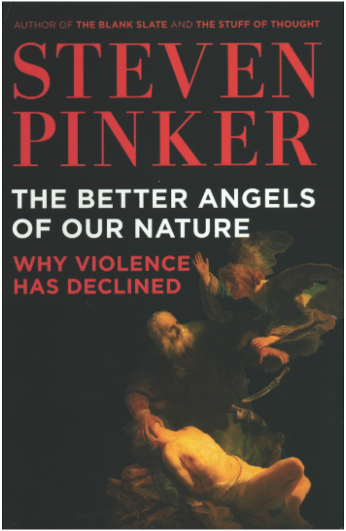
$$
\begin{aligned}
p(x = 5 \,|\, n = 10, p = .5) &= \binom{10}{5} (.5)^5 (1 - .5)^{10-5} \\
&= (252) * (.03125) * (.03125) \\
&= 0.2460938 \qquad (5)
\end{aligned}
$$

In R:

```
dbinom(5,10,.5)
```

```
## [1] 0.2460938
```

An Application: The Decline of War?



The Decline of War?

Pinker (2011) argues the absence of world wars since WW2 shows a decline of violence. But:

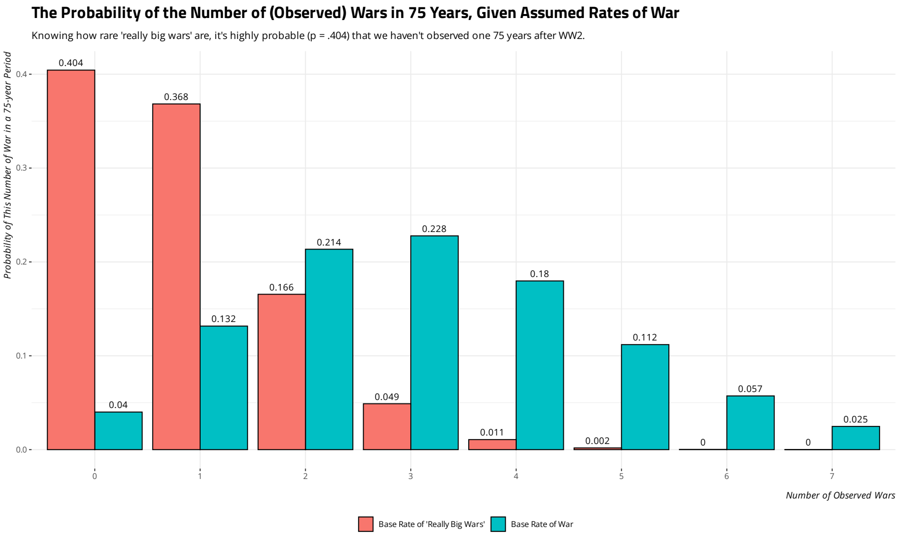
- This kind of war is fantastically rare.
- Gibler and Miller (Forthcoming) code 1,958 confrontations from 1816 to 2014.
- Of those: 84 are wars (p = .042)
- Of the wars, only 24 are wars we could think of as “really big” (p = .012)

The Decline of War?

The year is 2024. We haven't observed a World War II in, basically, 75 years. What is the probability of us *not* observing this where:

- p = .042, the overall base rate of war vs. not-war?
- p = .012, the overall base rate of a "really big war"?

```r
# library(tidyverse)
tibble(num_wars = c(0:7),
       base = dbinom(num_wars, 75, .042),
       rbw = dbinom(num_wars, 75, .012))
```


The Probability of the Number of (Observed) Wars in 75 Years, Given Assumed Rates of War
Knowing how rare 'really big wars' are, it's highly probable (p = .404) that we haven't observed one 75 years after WW2.
Probability of This Number of War in a 75-year Period
0.0
0.1
0.2
0.3
0.4
0.404
0.04
0.368
0.132
0.166
0.214
0.049
0.228
0.011
0.18
0.002
0.112
0
0.057
0
0.025
0
1
2
3
4
5
6
7
Number of Observed Wars
Base Rate of 'Really Big Wars'
Base Rate of War

Normal Functions

A “normal” function is also quite common.

- Data are distributed such that the majority cluster around some central tendency.
- More extreme cases occur less frequently.

Normal Density Function

We can model this with a **normal density function**.

- Sometimes called a Gaussian distribution in honor of Carl Friedrich Gauss, who discovered it.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \tag{6}$$

where: μ = the mean, σ^2 = the variance.

Normal Density Function

Properties of the normal density function.

- The tails are asymptote to 0.
- The kernel (inside the exponent) is a basic parabola.
  - The negative component flips the parabola downward.
- Denoted as a function in lieu of a probability because it is a continuous distribution.
- The distribution is perfectly symmetrical.
  - $-x$ is as far from μ as x.

Normal Density Function

x is unrestricted. It can be any value you want in the distribution.

- μ and σ^2 are parameters that define the shape of the distribution.
    - μ defines the central tendency.
    - σ^2 defines how short/wide the distribution is.

Demystifying the Normal Distribution

Notice: we're describing this distribution as a *function*. It does not communicate probabilities.

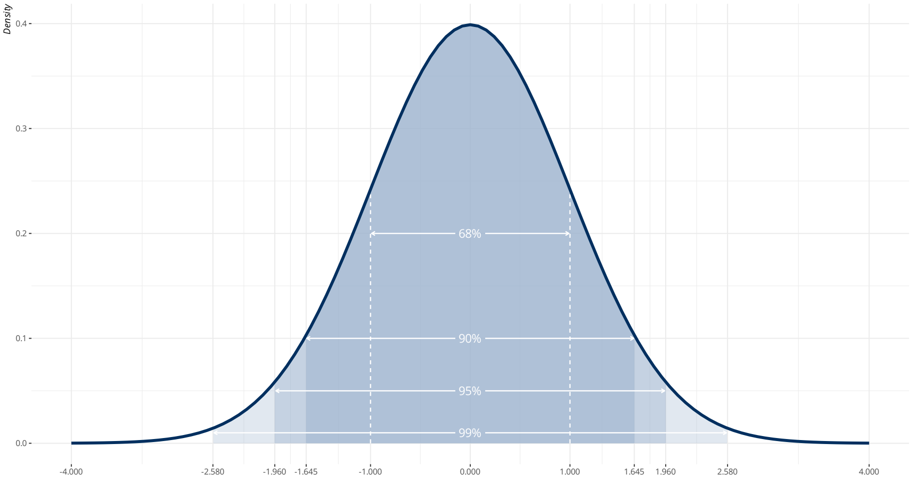
- The normal distribution is continuous. Thus, probability for any one value is basically 0.

That said, the area *under* the curve is the full domain and equals 1.

- The probability of selecting a number between two points on the x-axis equals the area under the curve *between* those two points.

The Area Underneath a Normal Distribution

The tails extend to infinity and are asymptote to zero, but the full domain sums to 1. 95% of all possible values are within about 1.96 standard units from the mean.



Central Limit Theorem

The **central limit theorem** says:

- with an infinite number samples of size *n*...
- from a population of *N* units...
- the sample means will be normally distributed.

Corollary findings:

- The mean of sample means would equal μ.
- Random sampling error would equal the standard error of the sample mean ($\frac{\sigma}{\sqrt{n}}$)

A Comment on Random Sampling

This theorem assumes a simple random sample of the population.

- Each unit has equal probability of inclusion vs. exclusion.

This introduces random sampling error, by definition.

- However, we can model this ($\frac{\sigma}{\sqrt{n}}$), and it's better than systematic sampling error.
- On the latter: read about the 1936 Literary Digest Poll.

An Applied Example from a Thermometer Rating

Let's use a real-world illustration from the 2020 ANES exploratory testing survey.

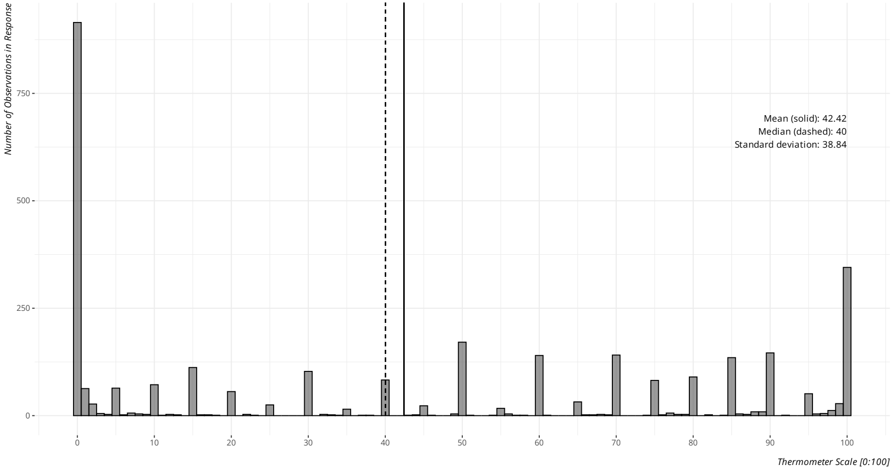
- Survey period: April 10-18, 2020 (online).
- Released July 27, 2020

The question is a basic thermometer rating of Donald Trump.

- Scale: 0 ("coldest") to 100 ("warmest")

Thermometer Ratings for Donald Trump (ANES ETS, 2020)

Thermometer ratings for divisive political figures in the U.S. tend to be ugly as hell with estimates of central tendency that don't faithfully capture the data.



Data: American National Election Studies (Exploratory Testing Survey, 2020). N = 3,073.

What We'll Do

Let's create a hypothetical "population" with the set parameters from the Trump ratings.

- Data will be bound between 0 and 100 with a mean of 42.42 and standard deviation of 38.84.
- N = 250,000 (i.e. scaled down from U.S. adult population of ~250 million).

We want to approximate the "population" mean thermometer rating via central limit theorem.

- We'll grab a million samples of ten respondents and store the sample means.

Let's plot the results.

R Code

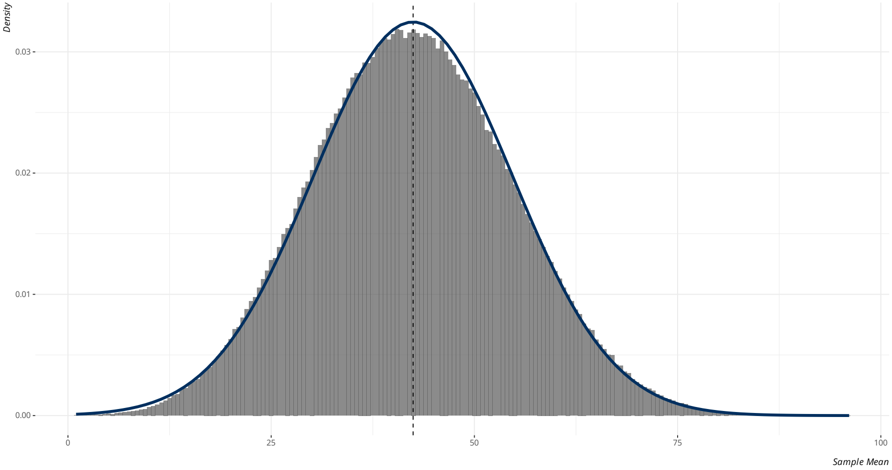
```r
# rbnorm() from {stevemisc}
Population <- rbnorm(250000, mean =42.42, sd = 38.84,
                     lowerbound = 0,
                     upperbound = 100,
                     round = TRUE,
                     seed = 8675309) # Jenny, I got your number...
```

Note: it's hard to perfectly mimic these kind of thermometer ratings from a simple distribution, but this will do.

- Mean: 42.459772
- Standard deviation: 38.8881803

The Distribution of 1,000,000 Sample Means, Each of Size 10

Notice the distribution is normal and the mean of sample means converges on the known population mean (vertical line).



Data: Simulated data for a population of 250,000 where mean = 42.42 and standard deviation = 38.84.

Implications of Central Limit Theorem

Infinite samples of *any* size (even absurdly small samples of high-variation data) reduce the gap between estimate and “true” population parameter.

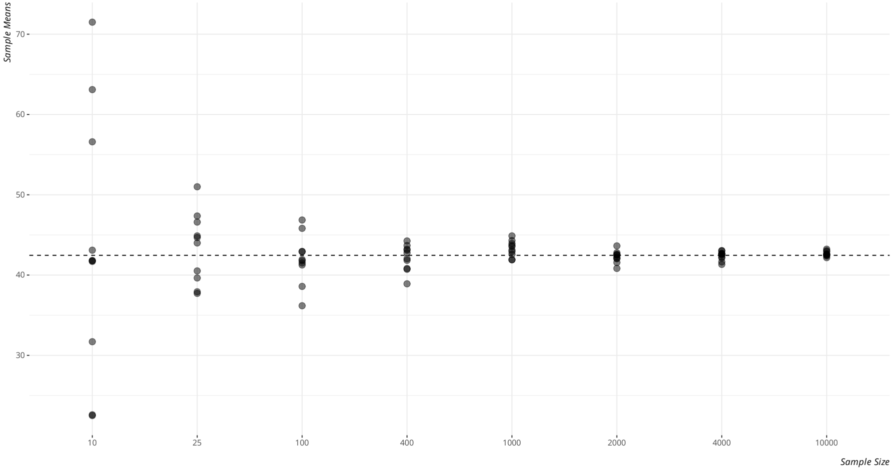
- Infinity samples gloss over the implications of random sampling error.

However, random sampling error decreases non-monotonically as a function of sample size.

- i.e. a good-sized sample reduces random sampling error in even high-variation data.

Ten Sample Means of Varying Sample Sizes from a Population

The diminishing returns of increasing sample size emerge around 1,000 observations, even as the spread in these simulated data is quite large.



Data: Simulated data for a population of 250,000 where mean = 42.42 and standard deviation = 38.84.

Conclusion

Probability (i.e. the stuff under the hood):

- Don't confuse $p(B|A)$ as equal to $p(A|B)$.
- Learn to "count".
- Understand some basic properties of a normal distribution.

Sampling:

- Central limit theorem: a population may not be "normal" but its random sampling distribution is.
- Random samples -> random sampling error (and that's fine).
- Absent infinity samples: get a good size sample.
- The sample statistic that emerges is the best guess of the population parameter.

Table of Contents