Preliminaries

Brad Jones[1]

[1] Department of Political Science
University of California, Davis

April 15, 2008

Today: Preliminaries

Preliminaries

- Today: Preliminary Concepts
- Most of which should seem familiar.
- If not, review regression text.

Setting the Stage: Basic Statistical Concepts

- Properties of probability
- Classic definition implies long-run relative frequency of some event A.
- Bayesians tell us, this is not always a good definition (they're right).
- However, let's walk before we run.
- $\Pr(A)$ is real-valued function defined on a sample space. Important properties:

$$0 \leq \Pr(A) \leq 1 \quad \forall \quad A \tag{1}$$
$$\Pr(A + B + C) = 1 \tag{2}$$
$$\Pr(A + B + C) = \Pr(A) + \Pr(B) + \Pr(C) \tag{3}$$

- (2) implies exhaustive events; (3) implies mutual exclusiveness.

Basic Concepts

- Random Variable: "a real-valued function defined on a sample space."
- It's an "observable" (with two flavors).
- Discrete and Continuous
- Probability Density Functions
- The PDF assigns probabilities to outcomes.

Discrete Random Variable

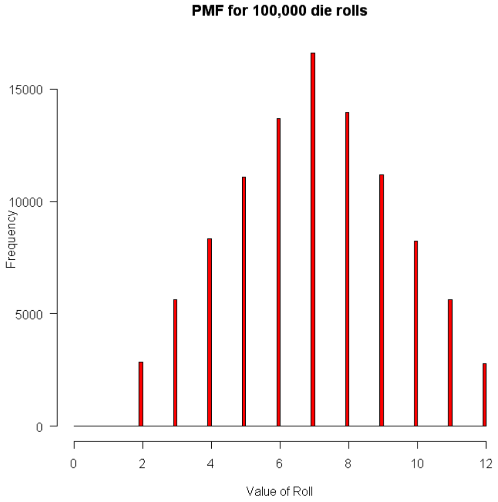
- PMF for a D.R.V.:

$$f(x) = \Pr(X = x_i) \quad \forall \quad i = 1, 2, \ldots, n \tag{4}$$
$$= 0 \quad \forall \quad x \neq x_i. \tag{5}$$

- In words: the probability that X is equal to some specific (discrete) value.
- Die Rolls in R:

```
d1 <- sample(1:6, 100000, prob=rep(1/6, 6),
replace=TRUE)
d2 <- sample(1:6, 100000, prob=rep(1/6, 6),
replace=TRUE)
die.roll <- d1 + d2
hist(die.roll, breaks= seq(0, 12, by=.1), prob=FALSE,
las=1, col="red", main="Distribution of 100,000 die
rolls", xlab = "Value of Roll")
```

Artwork

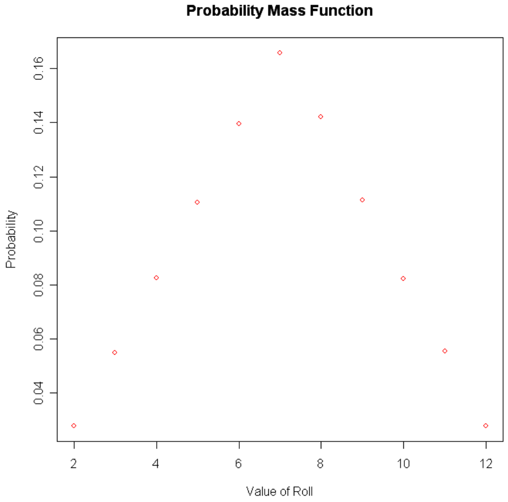

PMF for 100,000 die rolls
Frequency
0
5000
10000
15000
0
2
4
6
8
10
12
Value of Roll

Discrete Random Variable

- PMF in R:

```
freq<-cbind(table(die.roll))
prval<-freq/100000
roll<-cbind(2,3,4,5,6,7,8,9,10,11,12) plot(roll,
prval, col="red", main="Probability Mass Function",
xlab = "Value of Roll", ylab="Probability")
```

Artwork


Probability Mass Function
Probability
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
4
6
8
10
12
Value of Roll

Continuous Random Variable

- Continuous RVs have "density functions."
- The density is kind of like a "smoothed out" histogram.
- The probability of any *specific* realization of X is assumed to be 0. (Why?)
- $\therefore$ we must integrate to define probability (within an infinitesimally small differentiable area).
- $f(x)$ in discrete case is easy to define; in continuous case, $f(x)$ may take on a variety of forms.
- The PDF, $f(x)$, for the standard normal:

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{6}$$

- We use this distribution all the time: z-scores, for example.

Continuous Random Variable

- The cumulative distribution function obtains probabilities:

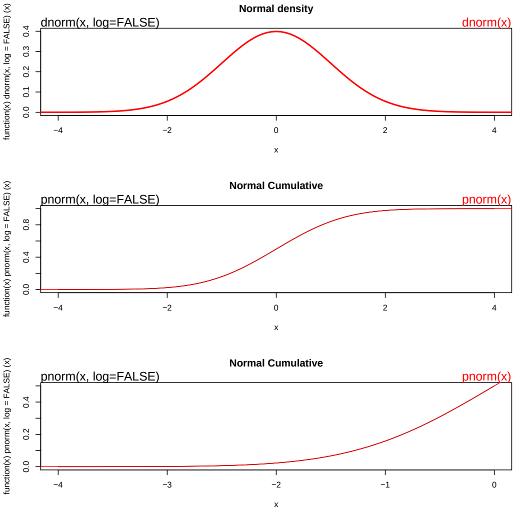
$$F(x) = \Pr(X \leq x) = \int_{x_{min}}^{x} f(x)d(x) \tag{7}$$

- Here, $f(x)$ is the PDF; $F(x)$ is the CDF.
- In a sense, the PDF is going to give us the "height" and the CDF gives us the area.
- Note that it must be the case all area under the curve must integrate to 1:

$$F(x) = \Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)d(x) = 1 \tag{8}$$

- Also important: $F(b) - F(a) = \Pr(a \leq X \leq b) = \int_a^b f(x)d(x)$ Remember this with ordinal logits and probits!

Artwork


Normal density
dnorm(x, log=FALSE)
dnorm(x)
function(x) dnorm(x, log = FALSE) (x)
x
Normal Cumulative
pnorm(x, log=FALSE)
pnorm(x)
function(x) pnorm(x, log = FALSE) (x)
x
Normal Cumulative
pnorm(x, log=FALSE)
pnorm(x)
function(x) pnorm(x, log = FALSE) (x)
x

Continuous Random Variable

- The classic linear model assumes *y* is continuous.
- It may not be (often, will not be).
- At "what point" regression fails us is a concern of this class.
- With a binary dependent variable, or categorical choice data, regression *will* fail us in certain kinds of ways (to certain degrees).
- But before we get to that, let's go on with a few more preliminaries.

Sampling Distributions

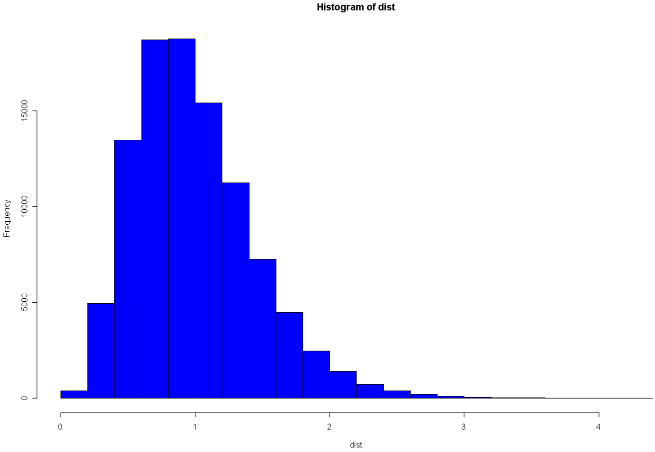
- The sampling distribution of a statistic is the probability distribution of a statistic obtained from repeated sampling.
- Suppose our statistic is θ.
- Let's review some basic sampling concepts (using R.
- Central Limit Theorem:
  The sum of many independently and identically distributed random variables will tend to a normal distribution in the limit. More specifically, if the sum of independently and identically distributed variables has a mean μ and a finite variance σ^2, then it will approximately follow a normal distribution.
- To sustain statistical inference, we rely heavily on the central limit theorem.
- Importantly, this result holds even if the population distribution is non-normal.

Let's create a world of 100,000 observations. This is our population and this is what it looks like:

```
> samplesize<-100000
>   dist<-sample(rgamma(samplesize,5,5))
>   hist(dist, col="blue1")
> meanX <- mean(dist); meanX
[1] 0.996132
```

Here, the mean of the distribution is 1.00. Call this μ. The population distribution looks like this:

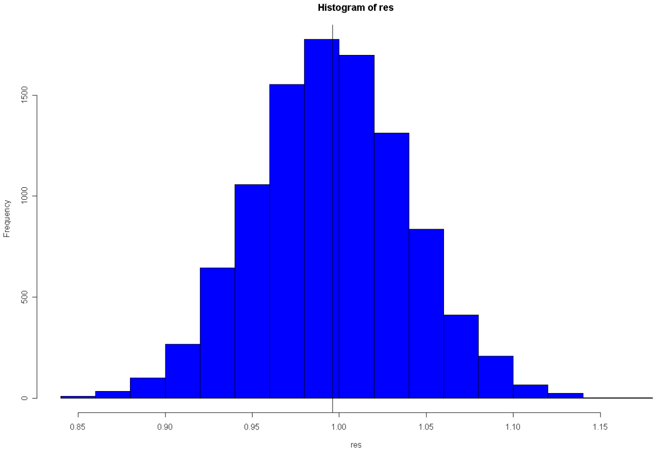
Artwork



The central limit theorem says that the distribution of “draws” of some statistic from the population, *even if the population is nonnormal*, will tend to a normal distribution. Suppose I took *one* sample of size 500?

```
> set.seed(510951)
> nsamp <- 1
> res <- numeric(nsamp)
> for (i in 1:nsamp) res[i] <- mean(sample(dist, 100, repla
> mean(res)
[1] 1.020171
```

The sample estimate, call it $\overline{X}$ is 1.02. It's “off” from μ. Suppose we were to take 10,000 samples of size 100?


Histogram of res
Frequency
0
500
1000
1500
0.85
0.90
0.95
1.00
1.05
1.10
1.15
res

Sampling Distribution

- This distribution is our sampling distribution.
- It gives the distribution of *means* from repeated samples.
- The mean of the sampling distribution is 1.00.
- The variability around the mean is the standard deviation.
- Inferentially, the standard error reported to you in regression output is "the standard deviation of the sampling distribution."
- The problem of estimation is we only have one sample with which to work.
- However, the nice thing about the CLT is "it gets us to the normal." (Or very close to it).

Properties of Estimators

- Estimators have no inherent use to us without properties.
- A random guess or mere dead reckoning *is* an estimator.
- It's just not very good.
- Let's review some properties of estimators.

Estimators

- $\hat{\theta}$ is what we're interested in.
- This is a statistic.
- Often, we're interested in the "first moment" of the sampling distribution.
- That is, the *mean of the sampling distribution.*
- The question is, what are the desirable properties of an estimator.

Small Sample Properties: Unbiasedness

- "Small sample properties."
- Unbiasedness: $E(\hat{\theta}) = \theta$.
- Equivalently: $E(\hat{\theta}) - \theta = 0$.
- The estimator is unbiased.
- In contrast: $E(\hat{\theta}) - \theta \neq 0$
- Implies biasedness in the estimator.
- Note: this property is a "repeated sampling" property.

Small Sample Properties: Efficiency

- Unbiasedness only tells us something about the central tendency of the sampling distribution.
- An estimator $\hat{\theta}$ is said to have "minimum variance" if var$(\hat{\theta}) < \mathrm{var}(\tilde{\theta})$.
- Efficiency: take two estimators, $\hat{\theta}$ and $\tilde{\theta}$. If each are unbiased *but* $\hat{\theta}$ is a minimum variance estimator, then $\hat{\theta}$ is efficient (or in words you may have used before, "best unbiased.").
- If $\hat{\theta}$ is a linear function of sample data, then $\hat{\theta}$ is a linear estimator.
- Thus, if $\hat{\theta}$ is efficient and linear, then in the class of linear estimators, it is "best unbiased."
- You of course have seen this: BLUE–best linear unbiased estimator.
- If the Gaussian assumptions hold, the OLS estimator has this property.

Large Sample Properties

- Some estimators will not satisfy these properties in small samples.
- Only in large samples do approximately equivalent properties hold.
- These kinds of properties are *asymptotic properties* or large sample properties.
- Asymptotic unbiasedness: $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$.
- Asymptotic properties are directly tied to sample sizes: small samples, they will not hold.
- You've seen this before:

$$s^2 = \frac{\sum(X_i - \overline{X})^2}{n}$$

- In small samples, this estimator for the variance is biased; in large samples, the bias tends to 0.
- Thus, the estimator is *asymptotically unbiased.*

Large Sample Properties: Consistency

- Consistency is a probabilistic statement:

$$\lim_{n \to \infty} \Pr\{\mid \hat{\theta} - \theta \mid < \delta\} = 1 \quad \delta > 0. \tag{9}$$

- Or $\text{plim}\hat{\theta} = \theta$.
- This is the consistency condition.
- Note that unbiasedness can hold for *any* sample size; consistency is purely an asymptotic property.
- A sufficient condition for consistency is that the bias and variance both tend toward 0 as n increases.
- Note that the *MSE* criteria is not used in the OLS context because it is biased. In large samples, this bias diminishes.
- The central limit theorem is an "asymptotic theorem."
- That is, asymptotic normality holds if the sampling distribution of $\hat{\theta} \to N$ as the sample size increases.

Toward Likelihood

- Know your estimator and its properties.
- Most all of the estimators from here on will *only* have large sample properties.
- It, therefore, is very risky to apply models considered from here onward, to small samples.
- That doesn't stop smart people from doing stupid things, however.
- Main point: research design, carefully conceived, is incredibly important.