14.30 Introduction to Statistical Methods in Economics

Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 15

Konrad Menzel

April 7, 2009

1 Special Distributions (continued)

1.1 The Poisson Distribution

Often, we may be interested in *how often* a certain event occurs in a given interval.

Example 1 *In airline safety, we may want to have some notion of how "safe" an airplane model is. The following data are from the website www.airsafe.com, and give the total number of flights and number of fatal events involving aircraft of a given type up to December 2006.*

Model	Flights	Events
Airbus A300	10.35M	9
Airbus A310	3.94M	6
Airbus A320/319/321	30.08M	7
Airbus A340	1.49M	0
Boeing 727	76.40M	48
Boeing 737	127.35M	64
Boeing 747	17.39M	28
Boeing 757	16.67M	7
Boeing 767	13.33M	6
Boeing 777	2.0M	0
Boeing DC9	61.69M	43
Boeing DC10	8.75M	15
Boeing MD11	1.69M	3
Boeing MD80/MD90	37.27M	14
Concorde	0.09M	1

We can see immediately that some types of aircraft have had fewer accidents than others simply because they haven't been in service for long or were only produced in small numbers. In order to be able to make a more meaningful comparison, we need a better way of describing the distribution of the number of fatal events.

Random variables of this type are commonly referred to as *count data*, and an often used distribution to describe them is the *Poisson* distribution

Definition 1 *X is said to follow* Poisson *distribution with* arrival rate *λ, $X \sim P(\lambda)$, if it has the p.d.f.*

$$f_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x \in \{0, 1, 2, \ldots\} \\ 0 & \text{otherwise} \end{cases}$$

Note that in particular, X is discrete.

Property 1 *For a Poisson random variable X,*

$$\begin{aligned} \mathbb{E}[X] &= \lambda \\ \text{Var}(X) &= \lambda \end{aligned}$$

In order to see why the Poisson distribution is a plausible candidate for the distribution of a count variable, let's do the following thought experiment: Suppose

- the probability that the event happens in a time window of length $\frac{1}{n}$ is $p_n = \frac{\lambda}{n}$.
- we also assume that the event happening at any given instant is independent across time.

We then let the partition of subintervals grow finer by letting n go to infinity. If the probability of two occurrences in the same shrinking subinterval goes to zero, and we then count the number of subintervals in which the event occurs at least once we get the total number of occurrences. Notice that this will be a binomial random variable with parameters $p = \frac{\lambda}{n}$ and n.

Proposition 1 *For the binomial random variable $X_n \sim B\left(n, \frac{\lambda}{n}\right)$, as $n \to \infty$, the p.d.f. converges to*

$$\lim_n f_{X_n}(x) = \lim_n \left(\begin{array}{c} n \\ x \end{array} \right) \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

PROOF: We can take the limit of the product as the product of limits, and evaluate each term separately: By a well-known result from calculus (e.g. can do Taylor expansions on both sides),

$$\lim_n \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

So we are left with the term

$$T_n = \left(\begin{array}{c} n \\ x \end{array} \right) \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{-x} = \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n - \lambda}\right)^x$$

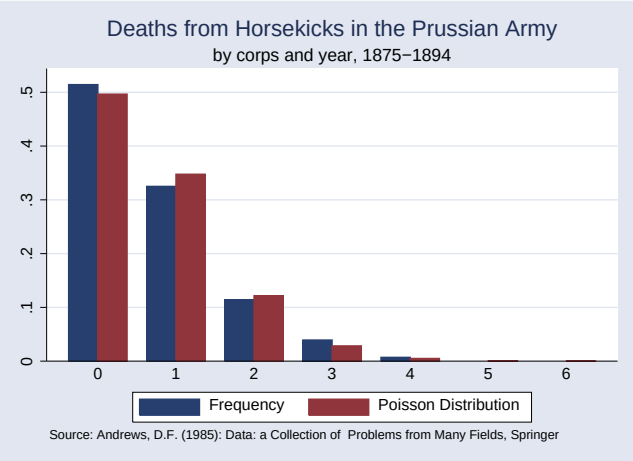
for which we can show

$$\lim_n T_n = \lim_n \frac{n(n-1)\ldots(n-k+1)\lambda^x}{x!(n-\lambda)^x} = \frac{\lambda^x}{x!} \lim_n \frac{n}{n-\lambda} \cdot \frac{n-1}{n-\lambda} \cdots \frac{n-x+1}{n-\lambda} = \frac{\lambda^x}{x!} \cdot 1$$

since both x and λ are fixed and therefore become small compared to n. Putting the pieces together, we get the expression stated in the proposition □

Example 2 *A classic example (classic in the history of statistics at least) for count data are records on the number of soldiers in the 19th century Prussian cavalry that died after being kicked by a horse. As found by the Russian statistician Ladislaus Bortkiewicz in 1898, the Poisson distribution gives a*

surprisingly good approximation to the observed frequencies of horsekick deaths over the course of a year in a given corps of the Prussian army.
So how do we compare the observed frequencies to a Poisson p.d.f. which, after all, depends on the unknown arrival rate λ? As a preview to our upcoming discussion of estimation later in this class, a plausible candidate for λ would be a value of the parameter in the p.d.f. which predicts the same expected *number of horsekick deaths that we observe in the sample. So for $X \sim P(\lambda)$, what is $\mathbb{E}[X]$? Well, as we argued above, a Poisson random variable is the limit of Binomial random variables $X_n \sim B\left(n, \frac{\lambda}{n}\right)$, where we let the number of trials n go to infinity. By our previous discussion of the Binomial distribution, $\mathbb{E}[X_n] = n\frac{\lambda}{n} = \lambda$, regardless of n. Hence, without directly evaluating the infinite series $\mathbb{E}[X] = \sum_{x=0}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!}$, we can say that $\mathbb{E}[X] = \lambda$. In the data set on horsekick deaths, the sample mean*



(by year and corps) is $\hat{\lambda} = 0.7$, and we can now plot the sample frequencies against the theoretical values of the Poisson p.d.f. for an arrival rate of $\lambda = 0.7$ in figure 2. The two distributions look strikingly similar, and this fact is often referred to as the "Law of Small Numbers."

2 Asymptotic Theory

Until now, we've assumed that we know the p.d.f. (or that we can find it), know parameters (such as μ and σ^2 for normal, λ for exponential etc.), and then make probability statements based on that knowledge.
In the next part of the class, we won't pretend that we have that information, but we'll construct functions of random variables - which are going to be the *estimators* - which, along with our knowledge of probability, will allow us to infer something about thee underlying distribution.
A particular estimator which plays an important rule in statistics is the *sample mean*, which typically approximates the expectation of a random variable in a sense we'll discuss in a few minutes.

Definition 2 *A* random sample *of size n is a sequence $X_1, \ldots, X_n$ of n random variables which are i.i.d., i.e. the X_i's are independent and have same p.d.f. $f_X(x)$.*

We also often refer to the *realizations* of the random variables as the random sample.
The main point of this lecture is going to be that if we have a random sample with n large, we actually don't need to know a lot about the distribution of X_i (e.g. don't need to know $f_{X_i}(x)$ in order to be able to describe the distribution of the sample mean fairly accurately.

The basic idea is that we'll find approximations to the p.d.f. which get closer and closer to the "truth" as we increase the size n of the sample. The two main results are:

1. the Law of Large Numbers: for large n, the sample mean will with all likelihood be "close" to the expectation $\mathbb{E}[X]$ of the random variable.
2. the Central Limit Theorem: the p.d.f. of the *standardized* sample mean ("standardized" in the sense of the last lecture: zero mean and unit variance) will become arbitrarily close to the p.d.f. of a standard normal random variable.

Formally, the asymptotic results state what will happen as $n \to \infty$, but for practical purposes (i.e. for finite n), they also imply that for n large enough, the approximations will be reasonably accurate.

2.1 The Law of Large Numbers

2.1.1 Chebyshev's Inequality

Chebyshev's Inequality is a formal result which gives a bound on the probability of the realization of a random variable being "far away" from the expectation.

Proposition 2 *Let X be a random variable with $\mathrm{Var}(X) < \infty$. Then for any $\varepsilon > 0$,*

$$P\left(|X - \mathbb{E}[X]| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}$$

PROOF: Let the p.d.f. of X be given by $f_X(x)$. We'll show that

$$\mathrm{Var}(X) \geq \varepsilon^2 P\left(|X - \mathbb{E}[X]| \geq \varepsilon\right)$$

By the definition of the variance,

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 f_X(t) dt \\
&= \int_{\mathbb{E}[X]-\varepsilon}^{\mathbb{E}[X]+\varepsilon} (t - \mathbb{E}[X])^2 f_X(t) dt + \int_{-\infty}^{\mathbb{E}[X]-\varepsilon} (t - \mathbb{E}[X])^2 f_X(t) dt + \int_{\mathbb{E}[X]+\varepsilon}^{\infty} (t - \mathbb{E}[X])^2 f_X(t) dt
\end{aligned}$$

Each of the three integrals is positive, and in addition, for any $t \leq \mathbb{E}[X] - \varepsilon$ or $t \geq \mathbb{E}[X] + \varepsilon$

$$(t - \mathbb{E}[X])^2 \geq \varepsilon^2$$

Therefore, we can just drop the first integral and get

$$\begin{aligned}
\mathrm{Var}(X) &\geq \int_{-\infty}^{\mathbb{E}[X]-\varepsilon} (t - \mathbb{E}[X])^2 f_X(t) dt + \int_{\mathbb{E}[X]+\varepsilon}^{\infty} (t - \mathbb{E}[X])^2 f_X(t) dt \\
&\geq \varepsilon^2 \int_{-\infty}^{\mathbb{E}[X]+\varepsilon} f_X(t) dt + \varepsilon^2 \int_{\mathbb{E}[X]+\varepsilon}^{\infty} f_X(t) dt \\
&= \varepsilon^2 P\left(|X - \mathbb{E}[X]| \geq \varepsilon\right)
\end{aligned}$$

Therefore, we can just divide through by ε^2 to get the result □

Remember that we said at an earlier point in this class that the variance of a random variable is a measure of its "dispersion." Chebyshev's Inequality makes this statement literal by relating the variance to the probability of observing "extreme" realizations (i.e. values that are far away from the mean) of the random variable X.

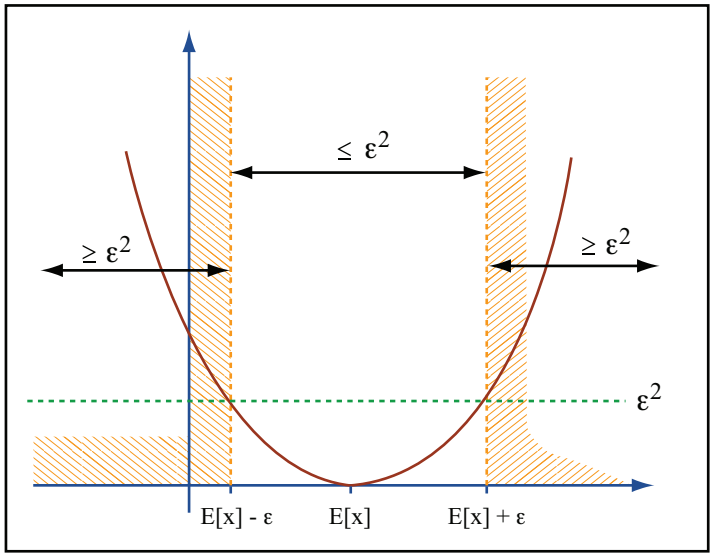


Image by MIT OpenCourseWare.

2.1.2 Law of Large Numbers

Definition 3 *The* sample mean *is the arithmetic average of n random variables (or realizations) from a random sample of size n. We denote it*

$$\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Note that since the X_is are random variables, $\bar{X}_n$ is also a random variable.

The expectation of the sample mean is

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_1]$$

If $X_1, \ldots, X_n$ are *independent*, the variance of the sample mean can be calculated as

$$\text{Var}\left(\bar{X}_n\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n}\text{Var}(X_1)$$

What if the X_is are i.i.d. normal, $X_i \sim N(\mu, \sigma^2)$? We know that a linear combination of normals is again normal with the appropriate variance and mean, so

$$\bar{X}_n \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$$

Since the variance decreases as we increase n, the mean will eventually be very close to $\mathbb{E}[X_i]$ with a very high probability. This is essentially what the Law of Large Numbers says:

Theorem 1 ***(Law of Large Numbers)*** *Suppose $X_1, \ldots, X_n$ is a sequence of i.i.d. draws with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 \infty$ for all i. Then for any $\varepsilon > 0$ (typically a small value), the sample mean satisfies*

$$\lim_n P\left(|\bar{X}_n - \mu| > \varepsilon\right) = 0$$

We say that $\bar{X}_n$ converges in probability *to* μ.

PROOF: We use our previous result that

$$\text{Var}(\bar{X}_n) = \frac{1}{n}\text{Var}(X_i) = \frac{\sigma^2}{n}$$

By Chebyshev's Inequality

$$P\left(|\bar{X}_n - \mu| > \varepsilon\right) \leq \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \overset{n\to\infty}{\longrightarrow} 0$$

This statement says that for large samples, the sample mean is unlikely to be far away from the expectation of the random variable. For a given n and a given variance σ^2, we can directly use Chebyshev's Inequality to bound the probability that the sample mean is more than a given distance away from μ.

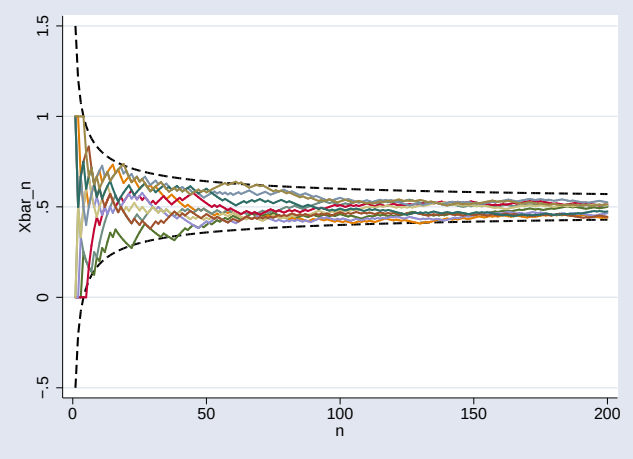


Figure 1: 10 sequences of average number of heads in n coin tosses - dashed lines are $\pm\frac{1}{\sqrt{2n}}$

Example 3 *Standardization of Units of Measurement (see Stigler's book): in the middle ages, often every city was using a different measure of "foot", "inch", "yard", etc. depending on how long the ruler's corresponding limb was. This means that there is a lot of variation in the units, which complicates trade and gives rise to many legal disputes, say whether a given bundle of cloth was really 20 yards long.*

One clever idea people came up with was the following: to determine the length of a rod of 16 feet, you take a random sample of 16 individuals (in this case the rule was the first 16 people to exit the church on Sunday morning), and add up the length of their feet to obtain a measurement of 16 feet, then divide that length by 16. See Figure 2. According to the formula for the variance of an average for 16 observations, this should decrease the variance of the new measurement unit by $\frac{1}{16}$.
If there were no systematic differences in foot sizes (or the church-going population) across different places, this should make it much easier for merchants from different places to trade with each other.

2.1.3 Example: The "Wisdom of Crowds"

Suppose that a population of size n chooses among 2 candidates for public office, where the candidate with a simple majority of the vote wins. We'll look at the random variable X_i which equals 1 if voter i

h! 

Figure 2: Woodcut by Köbel (1535) of 16 people determining the legal definition of a rod of 16 feet.

favors candidate A, and 0 otherwise. Candidate A wins if his vote share

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} \geq \frac{1}{2}$$

Candidate A is objectively the better choice, but this is only known to a proportion $2\varepsilon > 0$ who will vote for candidate A for sure, i.e. $P(X_i = 1) = 1$ for $i = 1, \ldots, 2n\varepsilon$. The remainder $1 - 2\varepsilon$ of the population doesn't have any substantial information about either candidates and casts votes at random with no preference for either candidate, i.e. $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ for $i = 2n\varepsilon, \ldots, n$. The vote share $\bar{X}_n$ for candidate A is given by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \varepsilon + \frac{1}{n}\sum_{i=2n\varepsilon}^{n} X_i$$

Therefore, the expected vote share for candidate A is

$$\mu = 2\varepsilon + \frac{1}{2}(1 - 2\varepsilon) = \frac{1}{2} + \varepsilon$$

and the variance is, by results for the binomial distribution,

$$\sigma^2 = (1 - 2\varepsilon)\frac{\frac{1}{2}\left(1 - \frac{1}{2}\right)}{n} = \frac{1 - 2\varepsilon}{4n}$$

By the same argument in the proof of Chebyshev's Inequality, we can start from the variance of $\bar{X}_n$ in order to derive bounds on the probabilities

$$\begin{aligned} \mathrm{Var}(\bar{X}_n) &= \int_{-\infty}^{\mu-\varepsilon} (t-\mu)^2 dt + \int_{\mu-\varepsilon}^{\mu+\varepsilon} (t-\mu)^2 dt + \int_{\mu+\varepsilon}^{\infty} (t-\mu)^2 dt \\ &\geq \varepsilon^2 \left[P(\bar{X}_n - \mu > \varepsilon) + P(\bar{X}_n - \mu < -\varepsilon)\right] \end{aligned}$$

Now note that since noise voters don't favor either candidate, the distribution is symmetric around μ, so that

$$P(\bar{X}_n - \mu > \varepsilon) = P(\bar{X}_n - \mu < -\varepsilon)$$

Since $\mu = \frac{1}{2} + \varepsilon$, the probability of candidate B losing can be obtained as

$$\text{Var}(\bar{X}_n) \geq 2\varepsilon^2 P(\bar{X}_n - \mu < -\varepsilon) \Leftrightarrow P\left(\bar{X}_n < \frac{1}{2}\right) \leq \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{1-2\varepsilon}{4n\varepsilon^2}$$

Let's try a few values: Say, $2\varepsilon = 5\%$, how large does n have to be in order to keep the probability of electing candidate B below 5%? - The bound becomes

$$P\left(\bar{X}_n < \frac{1}{2}\right) \leq \frac{90\%}{4(5\%)^2 n} = \frac{90}{n}$$

so $n \geq \frac{1800}{19} \approx 95$ is sufficient to keep the probability of electing for the wrong candidate below 5% even if 95% of the electorate take their decision at random.
This phenomenon is known as the "wisdom of the crowds:" the stochastic "noise" in the election outcome introduced by the uninformed voters averages out in large samples, and in the end only the systematic "signal" from the informed voters decides the outcome of the election.

Notice that, like in the Law of Large Numbers, we assumed that individuals' votes were independent. What happens if we drop independence?
Suppose that during the televised debate between the candidates, there is a fly in the TV studio which randomly lands on either A or B's face (with equal probability $\frac{1}{2}$) and spends some time crawling around, causing a lot of embarrassment to that candidate. The informed voters don't change their behavior, but the uninformed voters vote for the candidate with the fly with probability $\frac{1}{3}$, and with probability $\frac{2}{3}$ for his opponent.
By the law of iterated expectations, the mean of $\bar{X}_n$ is

$$\begin{aligned} \mu &= \mathbb{E}[\bar{X}_n|\text{fly lands on A}]\frac{1}{2} + \mathbb{E}[\bar{X}_n|\text{fly lands on B}]\frac{1}{2} \\ &= \frac{1}{2}\left(2\varepsilon + (1-2\varepsilon)\frac{1}{3}\right) + \frac{1}{2}\left(2\varepsilon + (1-2\varepsilon)\frac{2}{3}\right) = \frac{1}{2} + \varepsilon \end{aligned}$$

so the mean is the same as before. However, the variance does change: by the ANOVA identity (conditional variance)

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\mathbb{E}[\bar{X}_n|\text{fly}]\right) + \mathbb{E}\left[\text{Var}(\bar{X}_n|\text{fly})\right]$$

We can calculate

$$\begin{aligned} \text{Var}(\bar{X}_n|\text{fly on A}) &= \text{Var}(\bar{X}_n|\text{fly on B}) = \frac{2(1-2\varepsilon)}{9n} \\ \mathbb{E}[\bar{X}_n|\text{fly on A}] &= 2\varepsilon + (1-2\varepsilon)\frac{1}{3} = \frac{1}{3} + \frac{4}{3}\varepsilon \\ \mathbb{E}[\bar{X}_n|\text{fly on B}] &= 2\varepsilon + (1-2\varepsilon)\frac{2}{3} = \frac{2}{3} + \frac{2}{3}\varepsilon \end{aligned}$$

so that

$$\text{Var}(\bar{X}_n) = \left(\frac{1}{6} - \frac{1}{3}\varepsilon\right)^2 + \frac{2(1-2\varepsilon)}{9n}$$

Again, since the roles of the candidates can be exchanged for the noise voters, the distribution is symmetric around the mean, and we can use the bound on the probabilities derived above

$$P\left(\bar{X}_n < \frac{1}{2}\right) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2}$$

However, now we can see that $\mathrm{Var}(\bar{X}_n)$ doesn't go to zero anymore as $n \to \infty$ since the first term $\mathrm{Var}(\mathbb{E}[\bar{X}_n|\text{fly}])$ doesn't depend on n at all.
In numbers, if $\varepsilon = 15\%$ (six times larger than in the previous calculation), the bound equals

$$P\left(\bar{X}_n < \frac{1}{2}\right) \leq \frac{49}{81} + \frac{560}{81n}$$

so that the bound is far above $\frac{1}{2}$ no matter how large n is. Note that since this is only an upper bound, this doesn't really tell us how likely the event really is, but it is clear that since the variance doesn't decrease to zero, the "noise" voters will always have a strong influence on the election result.

Here the Law of Large numbers fails because the fly incident affects all "noise" voters at the same time, so $X_1, \ldots, X_n$ are no longer independent. The independence assumption is important because the reason why the law of large numbers usually works is that the "noise" averages out across many observations. If one component of the "noise" is common to (or at least highly correlated across) all observations, the variance contribution of this component - in our example the $\frac{49}{81}$ term in the bound on the probability - does *not* disappear even if the sample is very large.