



# Some Common Distributions

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# The Bernoulli distribution

- The **Bernoulli distribution** arises as the result of a binary outcome
- Bernoulli random variables take (only) the values 1 and 0 with probabilities of (say)  $p$  and  $1 - p$  respectively
- The PMF for a Bernoulli random variable  $X$  is

$$P(X = x) = p^x (1 - p)^{1-x}$$

- The mean of a Bernoulli random variable is  $p$  and the variance is  $p(1 - p)$
- If we let  $X$  be a Bernoulli random variable, it is typical to call  $X = 1$  as a "success" and  $X = 0$  as a "failure"

# Binomial trials

- The *binomial random variables* are obtained as the sum of iid Bernoulli trials
- In specific, let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ); then  $X = \sum_{i=1}^n X_i$  is a binomial random variable
- The binomial mass function is

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

for  $x = 0, \dots, n$

# Choose

- Recall that the notation

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

(read " $n$  choose  $x$ ") counts the number of ways of selecting  $x$  items out of  $n$  without replacement disregarding the order of the items

$$\binom{n}{0} = \binom{n}{n} = 1$$

# Example

- Suppose a friend has 8 children (oh my!), 7 of which are girls and none are twins
- If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

$$\binom{8}{7} .5^7 (1 - .5)^1 + \binom{8}{8} .5^8 (1 - .5)^0 \approx 0.04$$

```
choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
```

```
## [1] 0.03516
```

```
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
```

```
## [1] 0.03516
```

# The normal distribution

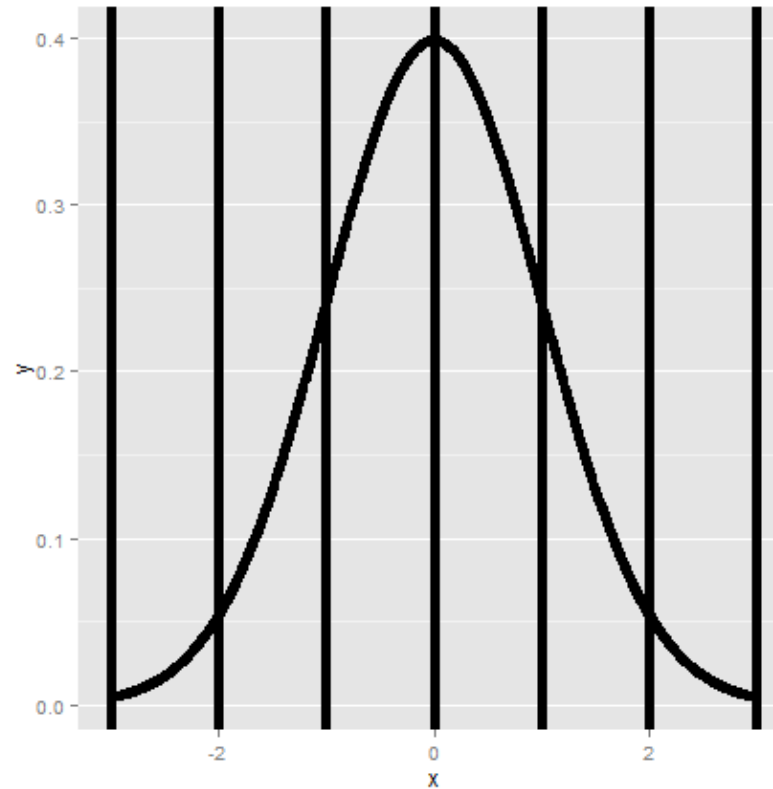
- A random variable is said to follow a **normal** or **Gaussian** distribution with mean  $\mu$  and variance  $\sigma^2$  if the associated density is

$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

If  $X$  a RV with this density then  $E[X] = \mu$  and  $Var(X) = \sigma^2$

- We write  $X \sim N(\mu, \sigma^2)$
- When  $\mu = 0$  and  $\sigma = 1$  the resulting distribution is called **the standard normal distribution**
- Standard normal RVs are often labeled  $Z$

# The standard normal distribution with reference lines



# Facts about the normal density

If  $X \sim N(\mu, \sigma^2)$  then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

If  $Z$  is standard normal

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$



# More facts about the normal density

1. Approximately 68%, 95% and 99% of the normal density lies within 1, 2 and 3 standard deviations from the mean, respectively
2.  $-1.28$ ,  $-1.645$ ,  $-1.96$  and  $-2.33$  are the  $10^{th}$ ,  $5^{th}$ ,  $2.5^{th}$  and  $1^{st}$  percentiles of the standard normal distribution respectively
3. By symmetry,  $1.28$ ,  $1.645$ ,  $1.96$  and  $2.33$  are the  $90^{th}$ ,  $95^{th}$ ,  $97.5^{th}$  and  $99^{th}$  percentiles of the standard normal distribution respectively

# Question

- What is the 95<sup>th</sup> percentile of a  $N(\mu, \sigma^2)$  distribution?
  - Quick answer in R `qnorm(.95, mean = mu, sd = sd)`
- Or, because you have the standard normal quantiles memorized and you know that 1.645 is the 95th percentile you know that the answer has to be

$$\mu + \sigma 1.645$$

- (In general  $\mu + \sigma z_0$  where  $z_0$  is the appropriate standard normal quantile)

# Question

- What is the probability that a  $N(\mu, \sigma^2)$  RV is larger than  $x$ ?

# Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What's the probability of getting more than 1,160 clicks in a day?

# Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What's the probability of getting more than 1,160 clicks in a day?

It's not very likely, 1,160 is 2.8 standard deviations from the mean

```
pnorm(1160, mean = 1020, sd = 50, lower.tail = FALSE)
```

```
## [1] 0.002555
```

```
pnorm(2.8, lower.tail = FALSE)
```

```
## [1] 0.002555
```

# Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What number of daily ad clicks would represent the one where 75% of days have fewer clicks (assuming days are independent and identically distributed)?

# Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What number of daily ad clicks would represent the one where 75% of days have fewer clicks (assuming days are independent and identically distributed)?

```
qnorm(0.75, mean = 1020, sd = 50)
```

```
## [1] 1054
```

# The Poisson distribution

- Used to model counts
- The Poisson mass function is

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for  $x = 0, 1, \dots$

- The mean of this distribution is  $\lambda$
- The variance of this distribution is  $\lambda$
- Notice that  $x$  ranges from 0 to  $\infty$



# Some uses for the Poisson distribution

- Modeling count data
- Modeling event-time or survival data
- Modeling contingency tables
- Approximating binomials when  $n$  is large and  $p$  is small

# Rates and Poisson random variables

- Poisson random variables are used to model rates
- $X \sim \text{Poisson}(\lambda t)$  where
  - $\lambda = E[X/t]$  is the expected count per unit of time
  - $t$  is the total monitoring time

# Example

The number of people that show up at a bus stop is Poisson with a mean of 2.5 per hour.

If watching the bus stop for 4 hours, what is the probability that 3 or fewer people show up for the whole time?

```
ppois(3, lambda = 2.5 * 4)
```

```
## [1] 0.01034
```

# Poisson approximation to the binomial

- When  $n$  is large and  $p$  is small the Poisson distribution is an accurate approximation to the binomial distribution
- Notation
  - $X \sim \text{Binomial}(n, p)$
  - $\lambda = np$
  - $n$  gets large
  - $p$  gets small

# Example, Poisson approximation to the binomial

We flip a coin with success probability 0.01 five hundred times.

What's the probability of 2 or fewer successes?

```
pbinom(2, size = 500, prob = 0.01)
```

```
## [1] 0.1234
```

```
ppois(2, lambda = 500 * 0.01)
```

```
## [1] 0.1247
```