Lecture 7: More Random Variables, and Intro to Estimation

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Misc

1. Announcement on Canvas
2. Show how to do tables in R Markdown
3. Recap

Random Variables: Example

Binomial RV

Based on several years of testing, it is determined that 96% of circuit boards are fully operational. A warehouse contains a very large population of boards. If 4 are selected at random, the distribution of $X =$ the number of operational boards in that sample of 4 would be described by a binomial RV.

Assumptions that must hold for $X \sim \text{Binomial}(n = 4, \pi = 0.96)$:

1. The "experiment" consists of 4 "sub-experiments" that each is a Bernoulli trial
    - the circuit board either works (1) or does not work (0)
2. The outcome of interest is the total number of successes in sub-experiments
3. The 4 "sub-experiments" are independent
    - By assumption
4. The probability of success for each "sub-experiment" is the same
    - each circuit board works with probability 0.96

Random Variables: Example

Binomial RV

Question: a customer orders 4 circuit boards for a critical job. In order to be able to complete the job, they need at least 3 of the circuit boards to be fully functioning. What is the probability that they can finish the job?

Recall: if $Y \sim \text{Binomial}(n, \pi)$, then $P(Y = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$.

By hand:

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) \\ &= \binom{4}{3} 0.96^3 \cdot (1 - 0.96)^1 + \binom{4}{4} 0.96^4 \cdot (1 - 0.96)^0 \\ &= 4 \cdot 0.885 \cdot 0.04 + 1 \cdot 0.849 \cdot 1 \\ &= 0.9906 \end{aligned}$$

Random Variables: Example

Binomial RV

Question: a customer orders 4 circuit boards for a critical job. In order to be able to complete the job, they need at least 3 of the circuit boards to be fully functioning. What is the probability that they can finish the job?

Recall: if $Y \sim \text{Binomial}(n, \pi)$, then $P(Y = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$.

In `R`: remember, $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2)$.

```
library(distributions3)
X <- Binomial(size = 4,
              p = 0.96)
```

```
1 - cdf(X, 2)
```

```
## [1] 0.9909043
```

```
sum(pmf(X, 3:4))
```

```
## [1] 0.9909043
```

Random Variables: Example

Binomial RV

The factory manufactures approximately 928 circuit boards a day. What it the expected number of working circuit boards manufactured in a day?

$X =$ number of working circuit boards in a day. Then $X \sim \text{Binomial}(n = 928, \pi = 0.96)$. So,

$$E(X) = n \cdot \pi = 928 \cdot 0.96 \approx 890.88.$$

What is the standard deviation of the number of working circuit boards manufactured in a day?

$$\text{Var}(X) = n \cdot \pi \cdot (1 - \pi) = 928 \cdot 0.96 \cdot 0.04 \approx 35.635$$

So

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{35.635} = 5.97.$$

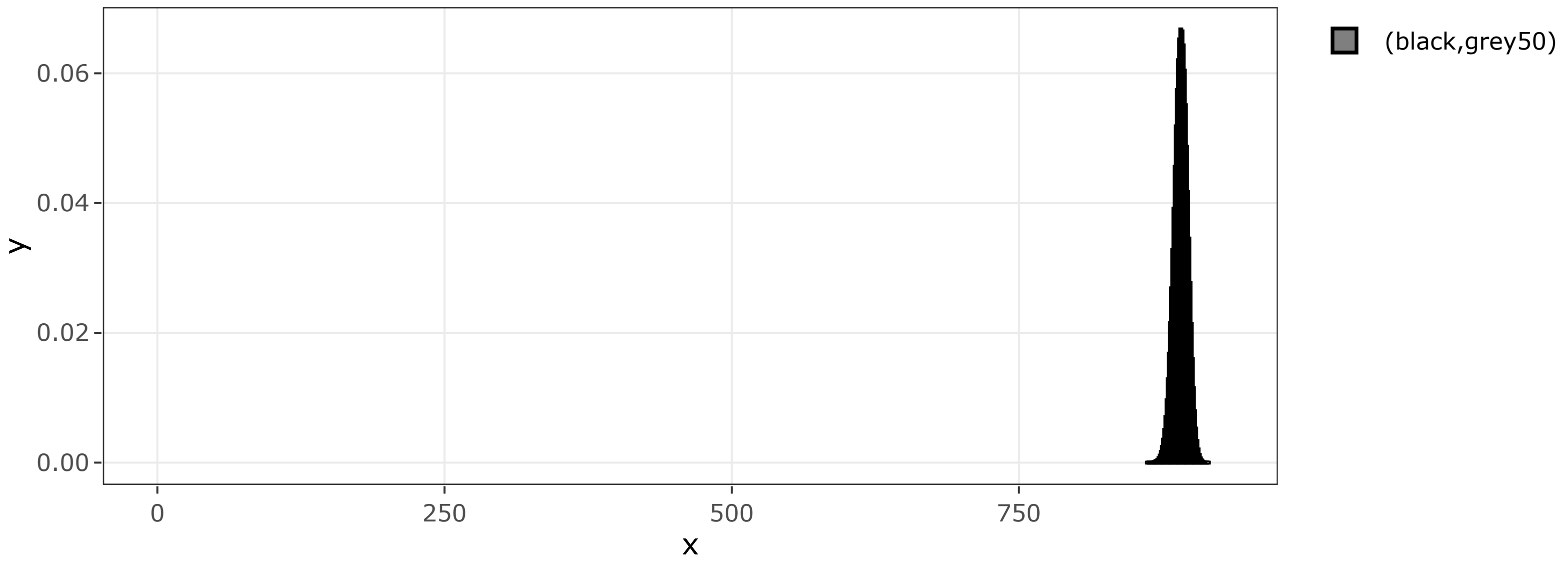
Random Variables: Example

Binomial RV

How many circuit boards can they with 98% certainty say will be working on any given day?

I.e. what is the x such that $P(X \geq x) = 0.98$?

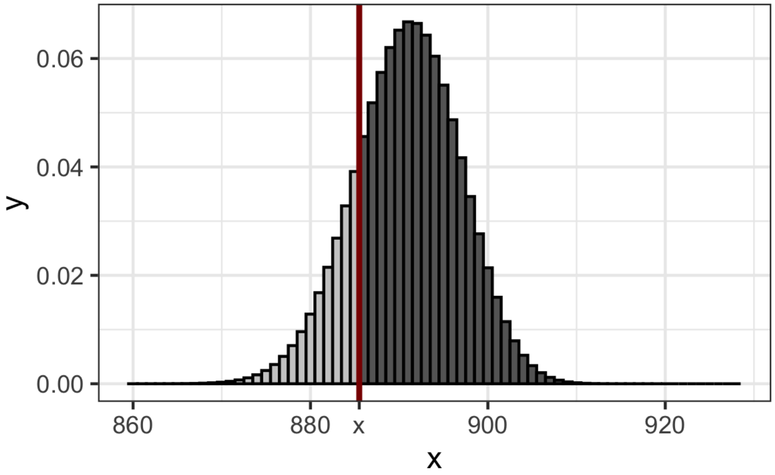
The p.m.f.:



Random Variables: Example

Binomial RV

I.e., we want to find x such that the area to the right is 0.98:



Or, similarly, find x so $P(X \leq x) = 1 - 0.98 = 0.02$. So we want to find the 2nd percentile. We do this using the quantile function:

```
X <- Binomial(928, 0.96)
quantile(X, 0.02)
```

```
## [1] 878
```

Random Variables: Example

Normal RV

Recall: previously, we almost died in a secret chamber. Fortunately, we survived, and decided to get really into ants.

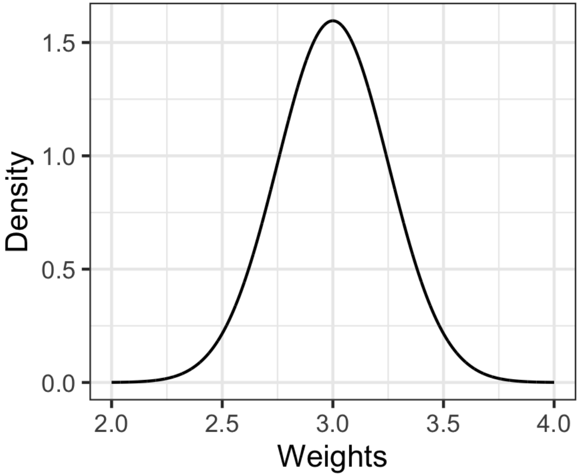
Somehow, we know that the weights of the ants in our ant farm follow a normal distribution with mean 3 mg and standard deviation 0.25 mg.

Weight of randomly chosen ant $= X \sim N(3, 0.25^2)$.

Random Variables: Example

Normal RV

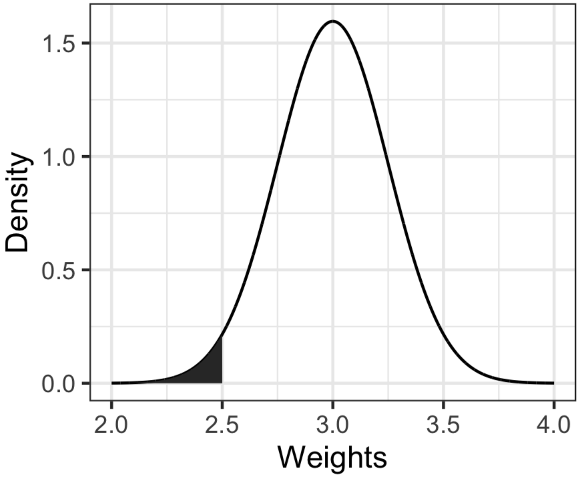
```
X <- Normal(mu = 3, sigma = 0.25)
```



Random Variables: Example

Normal RV

If we randomly select an ant, what is $P(X < 2.5)$?



```
cdf(X, 2.5)
```

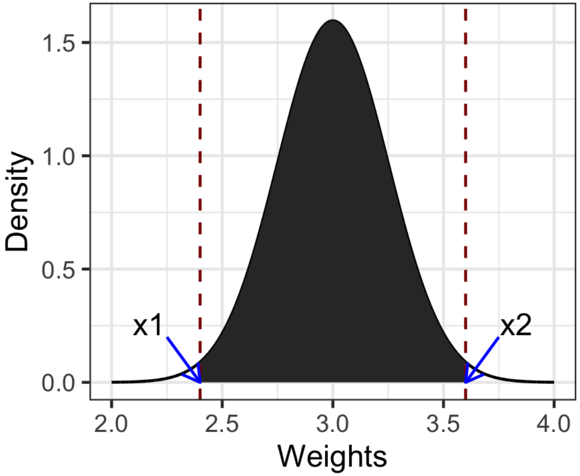
```
## [1] 0.02275013
```

Random Variables: Example

Normal RV

We know that ants are more likely to be closer to the expected value (i.e. closer to 3 mg). In what interval will the weights of most (say, 95%) of the ants be? Middle 95%. So, what are x_1 and x_2 such that

$$P(x_1 \leq X \leq x_2) = 0.95\%?$$

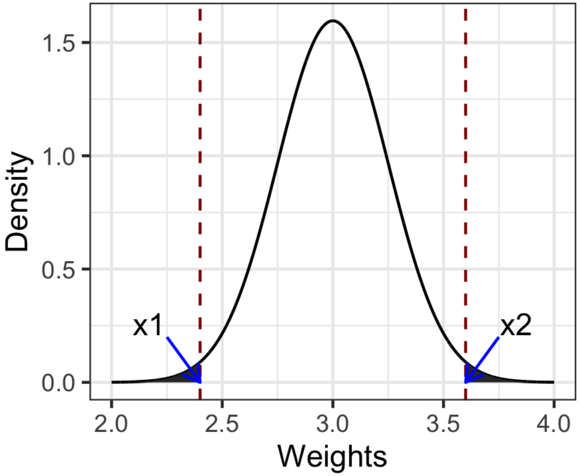


Random Variables: Example

Normal RV

We know that ants are more likely to be closer to the expected value (i.e. closer to 3 mg). In what interval will the weights of most (say, 95%) of the ants be? Middle 95%. So, what are x_1 and x_2 such that

$$P(X < x_1) = P(X > x_2) = 2.5\%?$$



```
quantile(X, 0.025)
```

```
## [1] 2.510009
```

```
quantile(X, 1-0.025)
```

```
## [1] 3.489991
```

Random Variables: Symmetrical Distributions

Definition: Symmetrical Distribution = the curve to the left of the mean is a mirror image of the curve to the right of the mean.

An important, and extremely useful, fact about symmetrical distributions: $P(X < \mu - x) = P(X > \mu + x)$.

In words: if you move the same distance to the left and right of the mean, the area to the left and right, respectively, is the same!

As a consequence, if $P(X < x_1) = P(X > x_2)$, then $|\mu - x_1| = |\mu - x_2|$.

In words: if the area (i.e. probability) to the left of one number is the same as the are to the right of another number, then the numbers are the same distance from the mean!

In particular, if we are considering the standard normal, i.e. $X \sim N(0, 1)$ (or any other symmetrical distribution with mean 0):

$$P(X < x_1) = P(X > x_2) \iff x_1 = -x_2.$$

Random Variables: Summary/Vocab

" $X \sim$ " = " X follows". This comes with:

- For discrete RV: Probability Mass Function (PMF) = $P(X = x)$
  - $P(X \text{ is something }) = \sum_x P(X = x) = 1$
- For continuous RV: Probability Density Function (PDF) = "the curve"
  - Total area under curve = 1
- PMF/PDF allows us to calculate probabilities for all possible events

Generally,

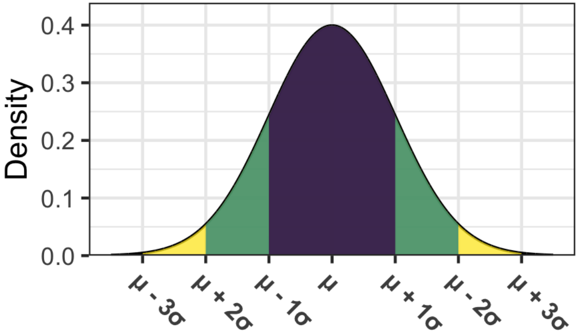
- Cumulative Density Function (CDF) = $P(X \leq x)$
  - "area under curve" to the left of x
- The q that satisfies $P(X \leq q) = p$ is called the p'th quantile
  - value that "cuts off" p to the left

Random Variables: Summary/Vocab

Normal Distribution:

- If $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ (X, Y independent)
  - $X \pm c \sim N(\mu_1 \pm c, \sigma_1^2), c \cdot X \sim N(c \cdot \mu_1, c^2 \cdot \sigma_1^2)$
  - $X \pm Y \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$

- $$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$$
- $$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$$
- $$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.998$$
- $$P(X < \mu - x) = P(X > \mu + x)$$
- $P(X < \mu) = P(X > \mu) = 0.5$



Random Variables: Summary/Vocab

Normal Distribution:

- We call $Z \sim N(0,1)$ the *standard normal*
- Standardization: $\frac{X - E(X)}{\text{SD}(X)} = \frac{X - \mu_1}{\sigma_1} \sim N(0,1)$

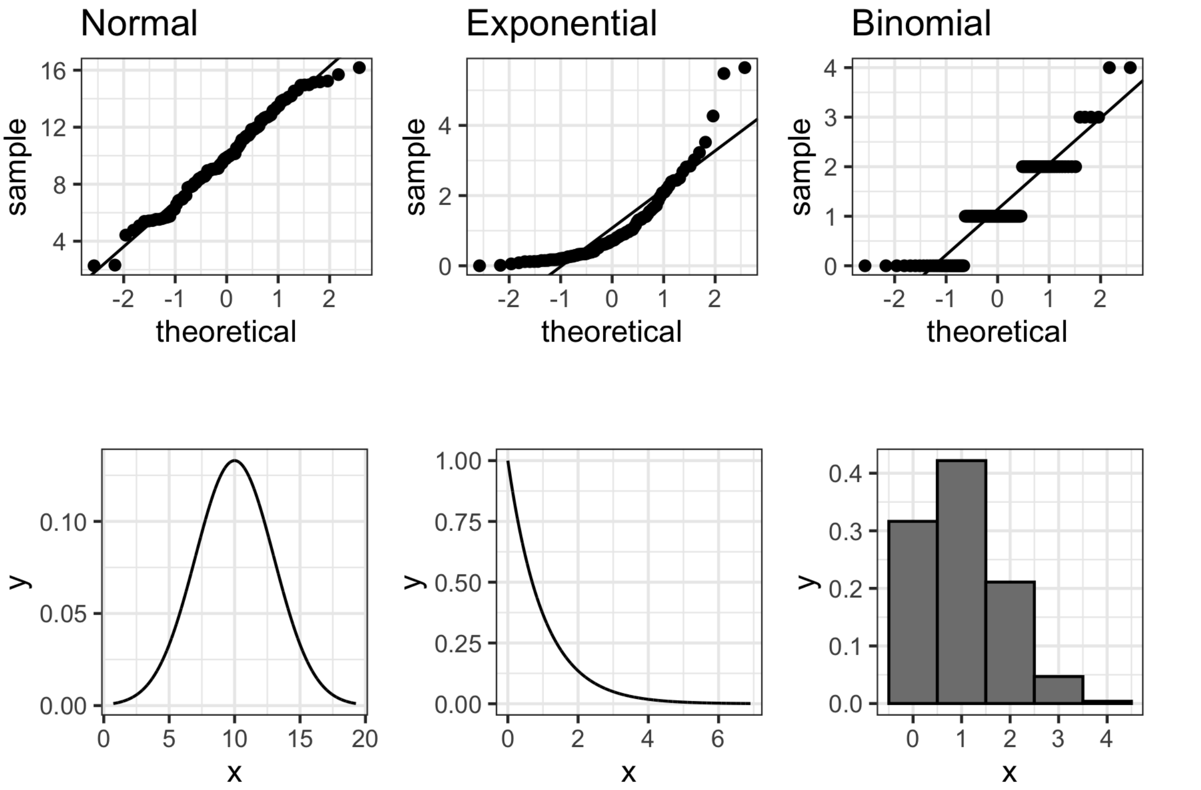
IMPORTANT NOTE: if you are ever asked to "standardize" something, it means "subtract the mean and divide by the standard deviation"!

Also sometimes called "z-value"/"t-value" - will talk much, much, much more about this later.

- The z that satisfies $P(X \geq z) = \alpha$ is called the α critical value
  - value that "cuts off" α to the right
  - denoted z_α.

Normal RV: One more thing

How to check for normality? QQ-plots.



Random Variables: Example

Normal RVs

Back to the ants: the weights of the ants in our ant farm follow a normal distribution with mean 3 mg and standard deviation 0.25 mg.

Weight of randomly chosen ant = $X \sim N(3, 0.25^2)$.

$P(2.75 < X < 3.25) = 0.68$ because $2.75 = \mu - \sigma$ and $3.25 = \mu + \sigma$.

$P(2.5 < X < 3.5) = 0.95$ because $2.5 = \mu - 2\sigma$ and $3.5 = \mu + 2\sigma$.

Estimation

The art of coming up with our "best guess" for the truth. This is called *an estimate*.

An *estimator* is a function that takes values from a sample and provides an estimate ("best guess").

Estimation

In order to come up with a good estimator, it is important to know how the sample was gathered. Three important definitions:

- a sample is called a **simple random sample** (SRS) if every possible element is equally likely to be sampled
    - unless otherwise stated, **all samples in this class are SRS**
- a sample is drawn **with replacement** if an element is replaced to the population before the next element is drawn. Otherwise, we say it is drawn **without replacement**.
    - without replacement = each element can only be sampled once.
- a collection of RVs $X_1, X_2, \ldots, X_n$ are said to be **independent and identically distributed** (iid) if
    1. they are all independent of each other
    2. they all follow the same distribution.

Technically, SRS **ONLY** if done with replacement. However, if population is "big enough", sample without repleacement is approximately the same as with replacement. (Recall last weeks discussion.)

Estimation

Estimating Population Mean

- A car manufacturer uses an automatic device to apply paint to engine blocks.
- engine blocks get very hot, so the paint must be heat-resistant,
- important that the amount applied is of a minimum thickness
- warehouse contains thousands of blocks painted by the automatic device
- he manufacturer wants to know the average amount of paint applied by the device

16 blocks will be selected at random, and the paint thickness measured in mm. Let $X_1, \ldots, X_{16}$ be RVs indicating the thickness of the 16 blocks.

Let's assume these RVs are iid -- i.e. independent and identically distributed. There exists some true expected value of these: $E(X_i) = \mu$. There also exists some true variance, $\mathrm{Var}(X_i) = \sigma^2$.

Estimation

Estimating Population Mean

Now we actually observe 16 realizations of these RVs:

```
paint_thickness <- data.frame(
  thickness = c(1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83,
                1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49)
)
```

Using these data, what would be your "best guess" for the true mean μ?

I would use the sample average:

```
paint_thickness %>%
  summarize(Mean = mean(thickness))
```

```
##       Mean
## 1 1.348125
```

This **estimATE** comes from the **estimatOR**: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Notice: the **estimator** is a **RV** while the **estimate** is a **realization** of that RV.