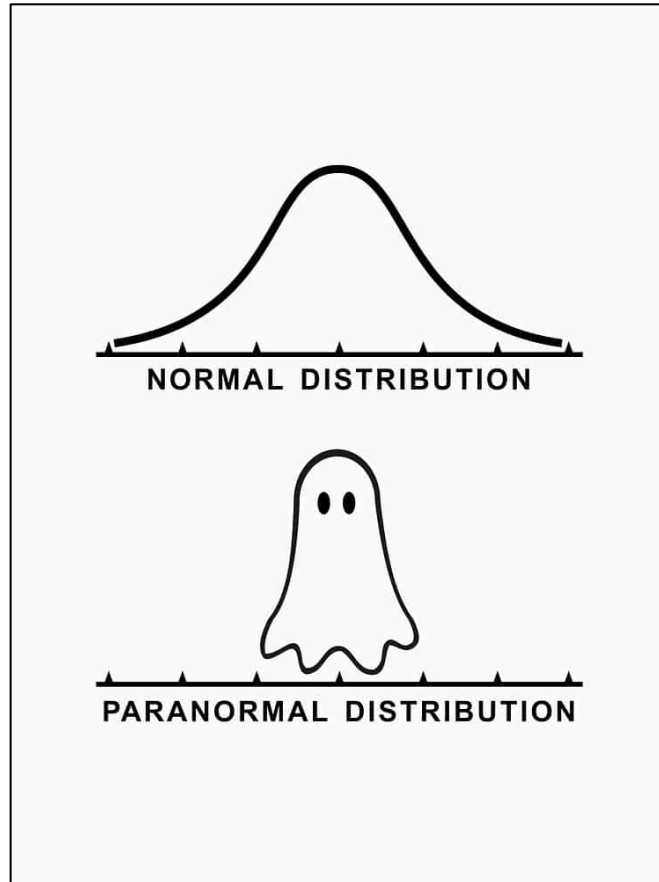


Variables, Distributions, Sampling

ENTMLGY 6707 Entomological Techniques and Data Analysis



Learning objectives

1. Differentiate between categorical and continuous variables
2. Compare common probability distributions – and their parameters - used to model biological data
3. Describe basic parameters and equations for summarizing data

Variables

Variables are properties with respect to which individuals in a sample differ in some ascertainable way. If properties do not differ (= no variability), they cannot be of statistical interest.

Variables come in two general forms:

Continuous: numeric, can range from $-\infty$ to ∞ ; "measured"

Examples: Time (seconds, minutes), height (m), weight (g), elevation, temperature ($^{\circ}\text{C}$), precipitation (mm)

Categorical: also called discrete or discontinuous; has two or more categories; "counted"

Categorical variables

Nominal: categories without a number assigned to them.

Examples: biological sex; species; survival; survey data (yes/no)

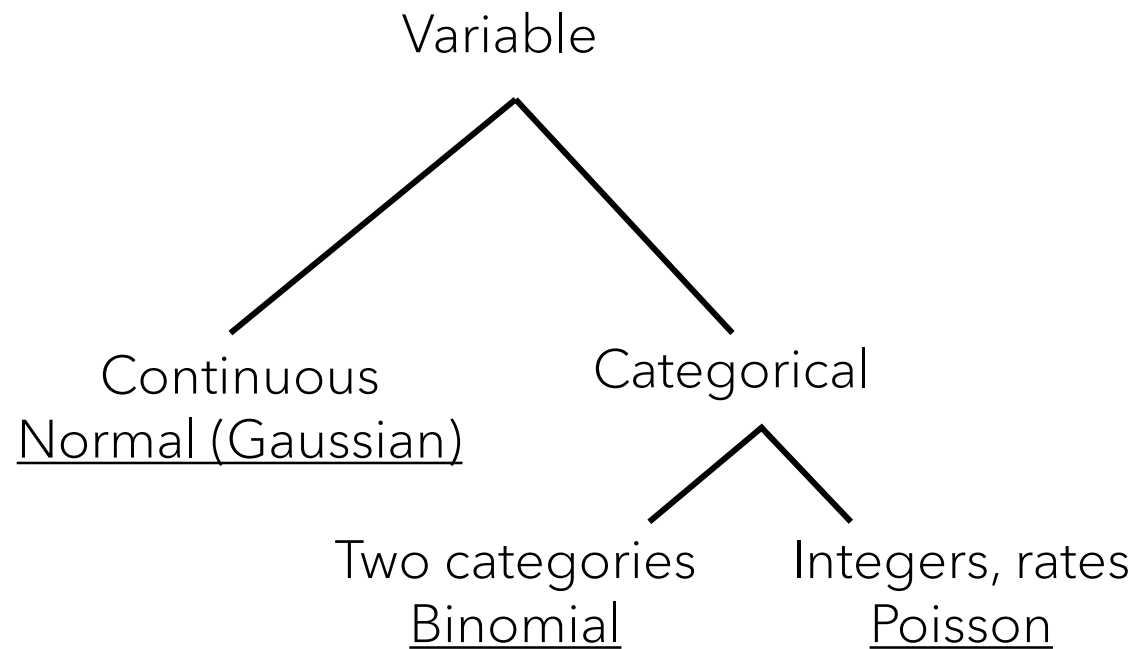
Count data: integers, rates (= integers per unit of time or space)

Examples: insects or plants per m²

Ordinal: categories in ordered form

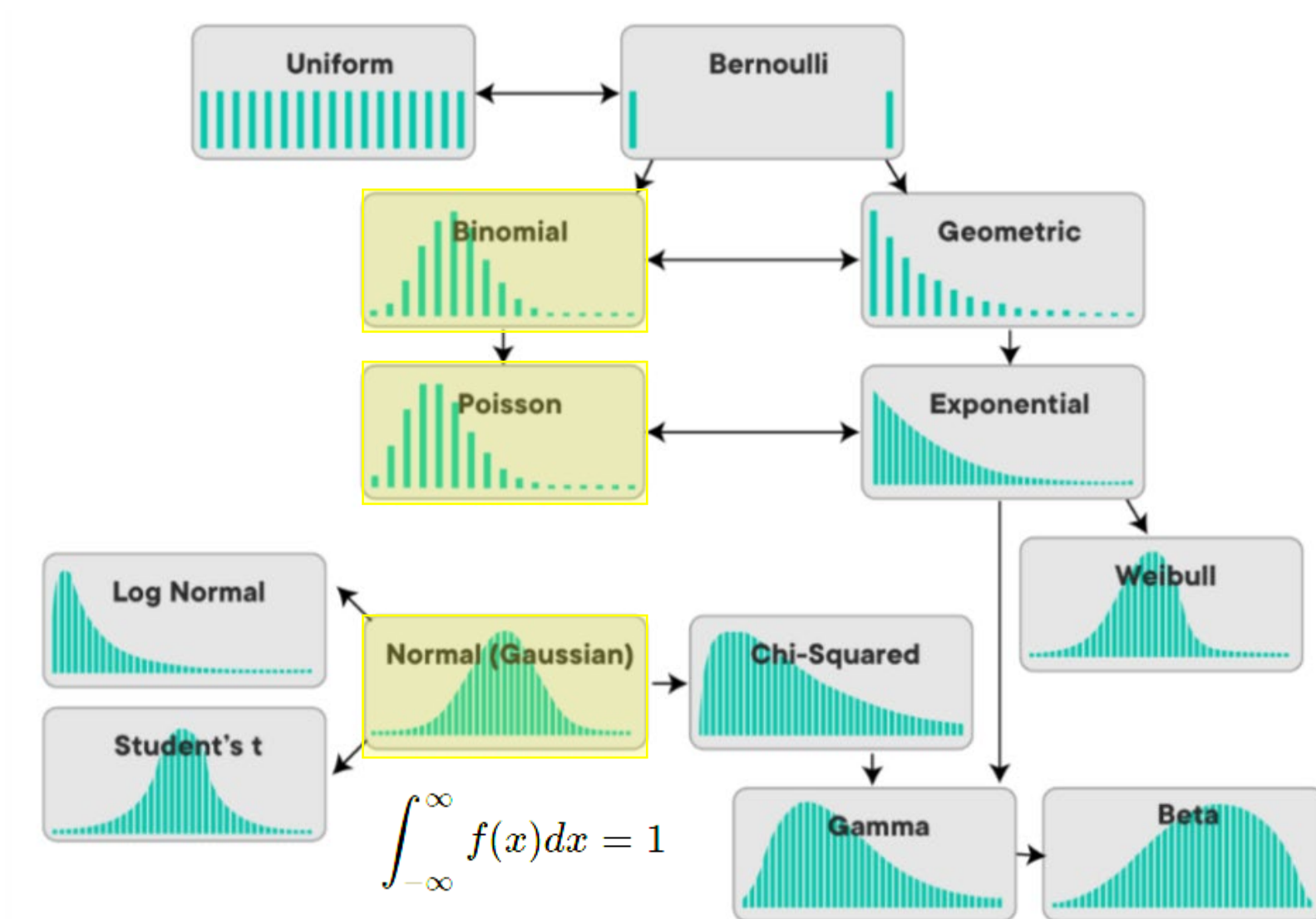
Examples: survey data (disagree, neutral, agree); crown class in forestry (co-dominant, dominant intermediate, suppressed)

For a given variable, what's the expected distribution?

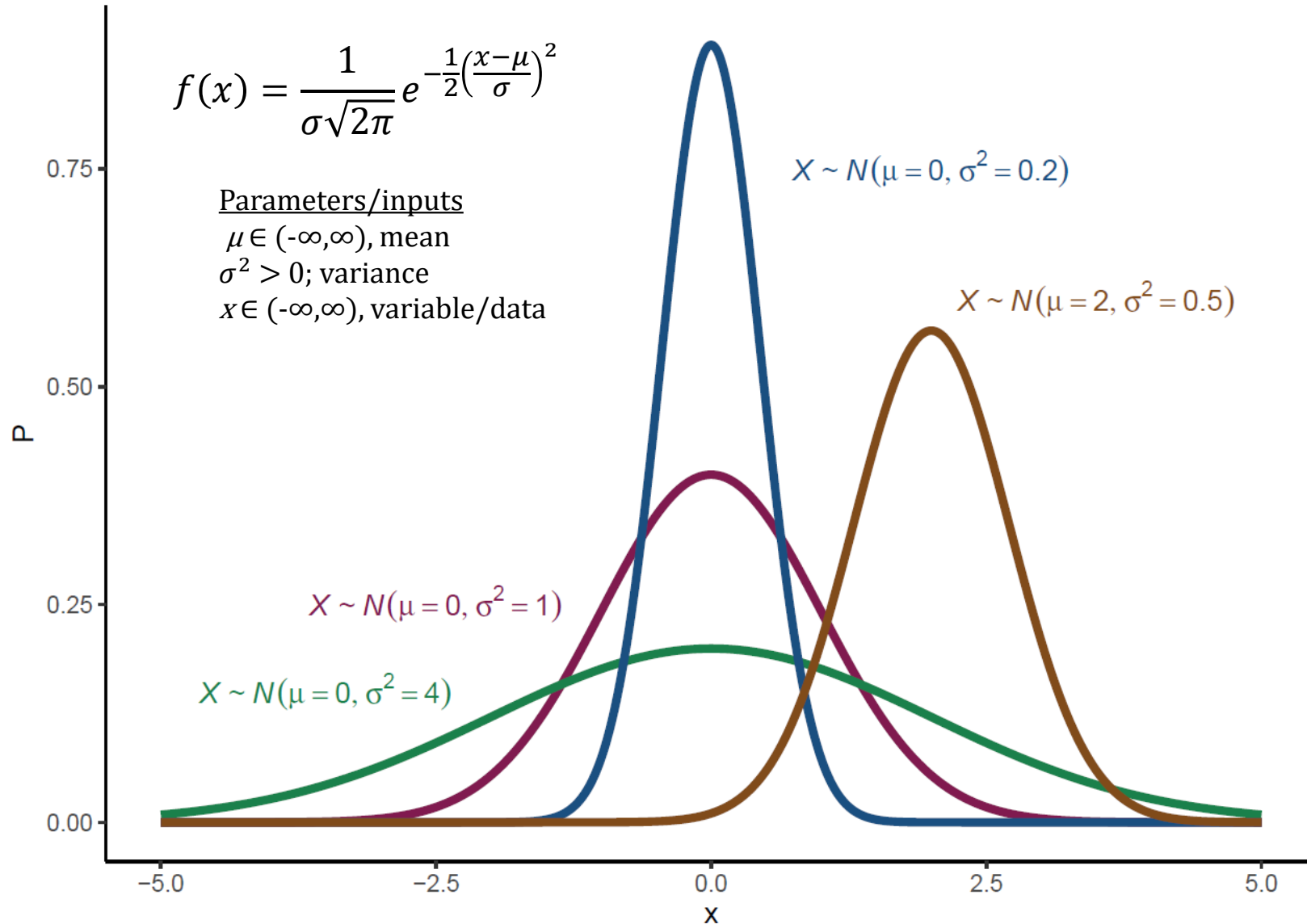


Activity (**submit to "In class 4" on Carmen**) : Describe one **response** variable and one **predictor** you are measuring as part of your thesis. Identify (i) whether each one is continuous vs. categorical, (ii) the expected distribution of the response variable, (iii) what analysis (e.g., ANOVA, simple linear regression, logistic regression, etc.) you plan to use.

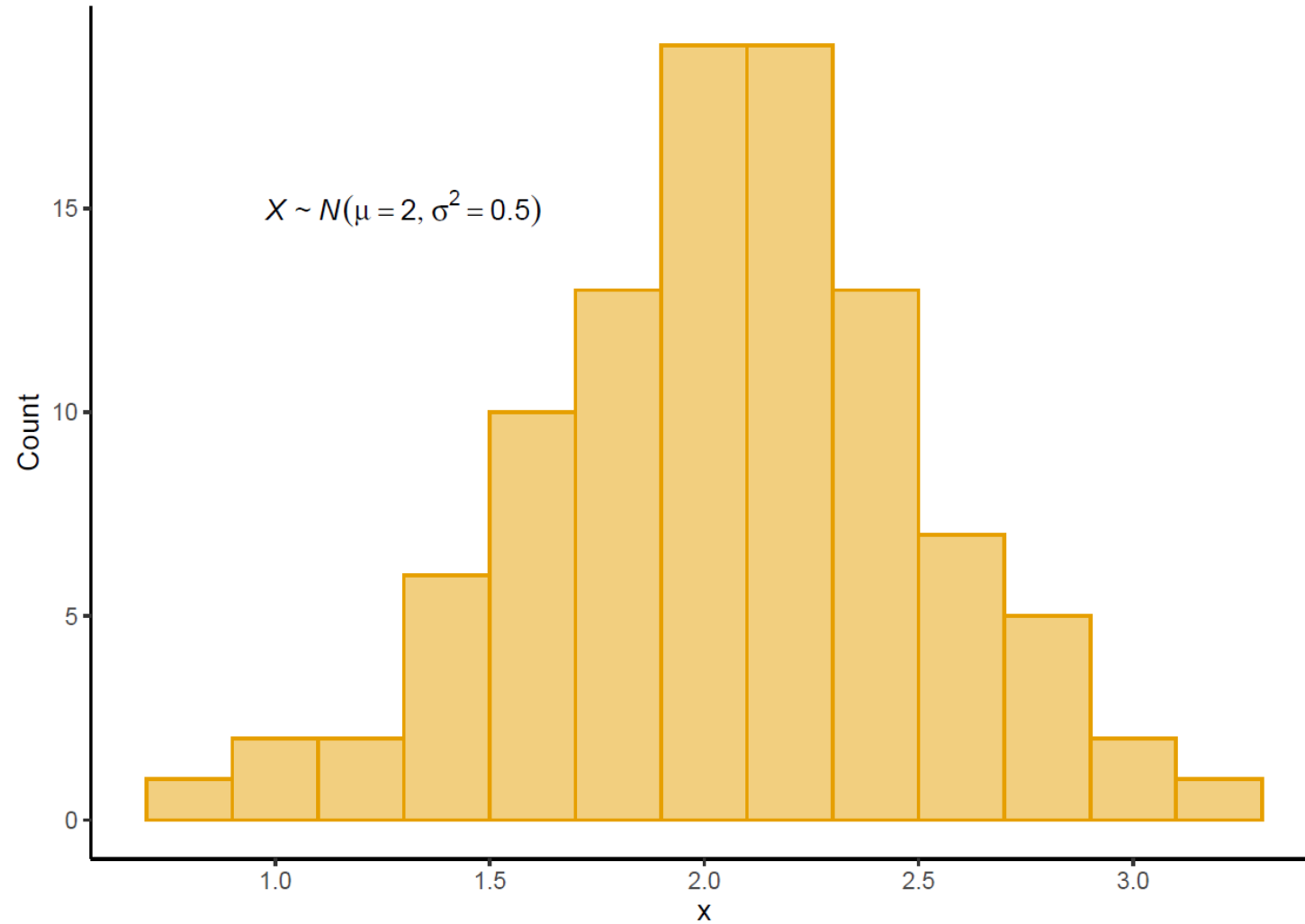
Probability distributions...there's a lot of 'em



Normal (Gaussian) distribution



Normal (Gaussian): sample



Normal (Gaussian): example



Placed caterpillars individually onto one of two plant species, measure larval size (head capsule width) in the final instar.

| plant_id | size_mm | plant_species |
|----------|---------|---------------|
| 1 | 1.3 | A |
| 2 | 1.2 | A |
| 3 | 1.3 | A |
| 4 | 1.4 | B |
| 5 | 1.6 | B |
| 6 | 1.5 | B |
| | | |

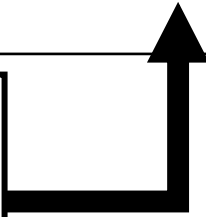
```
fit1 <- lm(size_mm~plant_species, data=df)
```

Mean, variance, SE

Activity: I have provided (i) some of the answers and (ii) formulas for the last column (in random order). Please try to fill out the rest from memory.

| Entity | Population parameter | Sample statistic | Estimating formula |
|------------------------|----------------------|------------------|--------------------|
| Value for member | X_i | x_i | Empty |
| No. of members | N | | Empty |
| Arithmetic avg (mean) | | | |
| Variance | | | |
| Standard deviation | σ | s | |
| Standard error of mean | σ_μ | | |

| | | | |
|--|--------------|------------------------|------------------------|
| a | b | c | d |
| $\frac{\sum (x_i - \bar{x})^2}{n - 1}$ | $\sqrt{s^2}$ | $\sqrt{\frac{s^2}{n}}$ | $\frac{1}{n} \sum x_i$ |



Mean, variance, SE

| Entity | Population parameter | Sample statistic | Estimating formula |
|------------------------|----------------------|------------------|--------------------|
| Value for member | X_i | x_i | . |
| No. of members | N | | . |
| Arithmetic avg (mean) | | | d |
| Variance | | | a |
| Standard deviation | σ | s | b |
| Standard error of mean | σ_μ | | c |

a

b

c

d

$$\frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$$\sqrt{s^2}$$

$$\sqrt{\frac{s^2}{n}}$$

$$\frac{1}{n} \sum x_i$$

Mean, variance, SE

| Entity | Population parameter | Sample statistic | Estimating formula |
|------------------------|----------------------|------------------|--|
| Value for member | X_i | x_i | . |
| No. of members | N | n | . |
| Arithmetic avg (mean) | μ | \bar{x} | $\frac{1}{n} \sum x_i$ |
| Variance | σ^2 | s^2 | $\frac{\sum (x_i - \bar{x})^2}{n - 1}$ |
| Standard deviation | σ | s | $\sqrt{s^2}$ |
| Standard error of mean | σ_μ | $SE_{\bar{x}}$ | $\sqrt{\frac{s^2}{n}}$ |

Binomial distribution

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

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

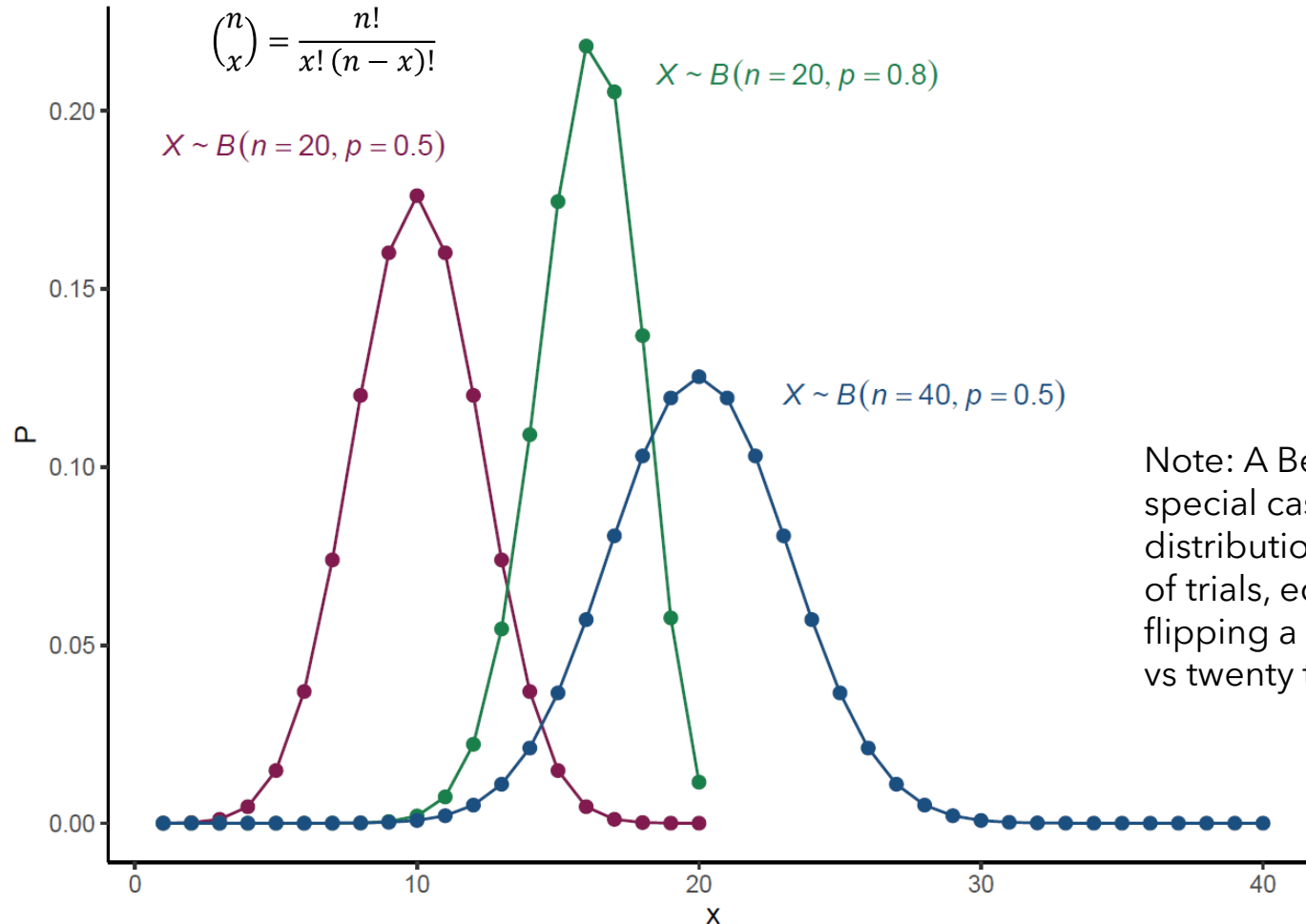
Parameters/inputs

$n \in (0, \infty)$, number of trials

$p \in (0, 1)$, probability of success per trial

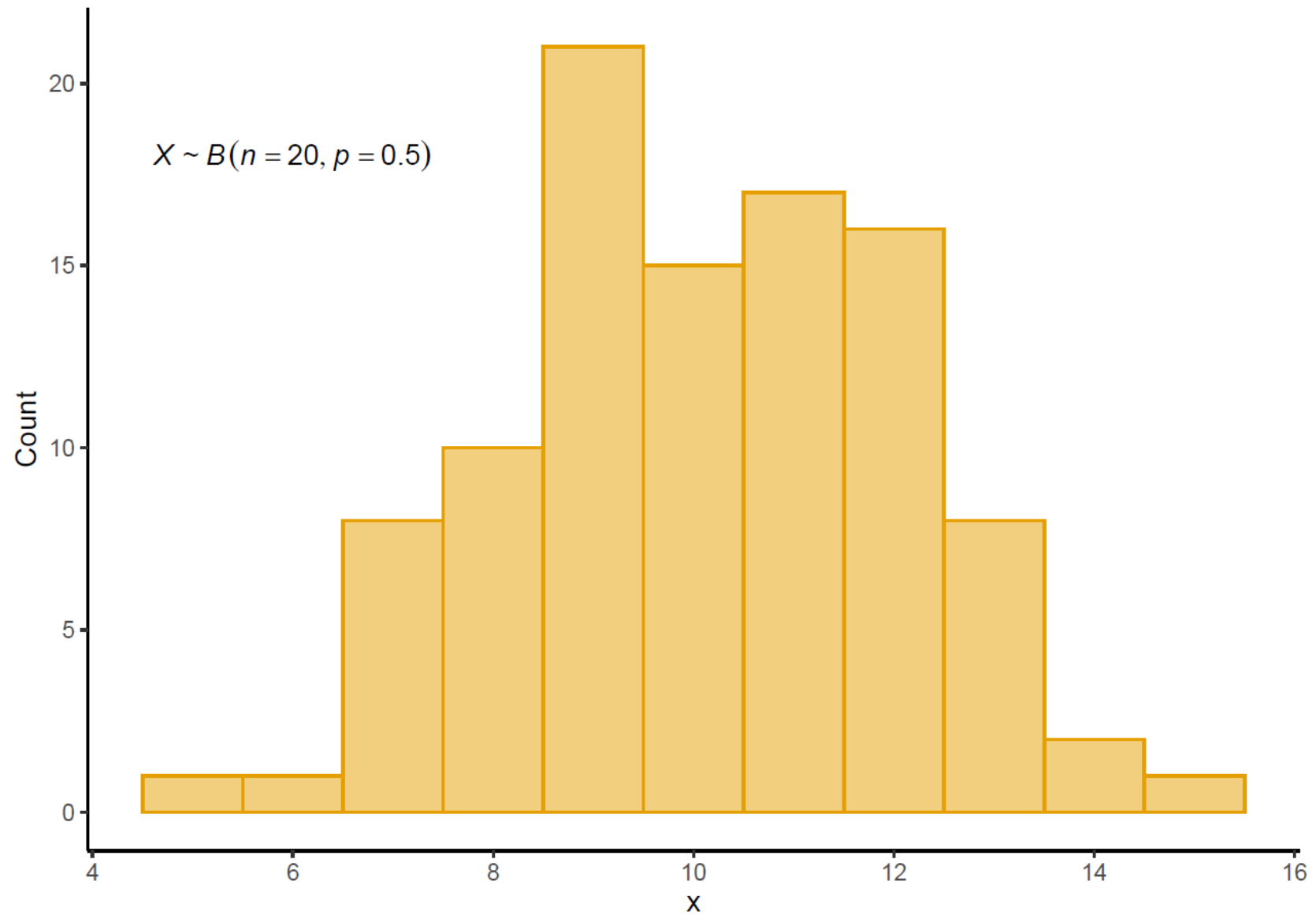
$q = 1 - p$; probability of failure per trial

$x \in (0, n)$, number of times a given outcome occurred across n trials



Note: A Bernoulli distribution is a special case of a Binomial distribution in which n , the number of trials, equals 1. For example, flipping a coin once (= Bernoulli) vs twenty times (= Binomial)

Binomial distribution



Summary of distributions and their parameters

| Distribution | Mean | Variance | SE |
|--------------|-------|-----------------|-----------------------------|
| Normal | μ | σ^2 | $\sqrt{\frac{\sigma^2}{n}}$ |
| Binomial | np | $np(1-p) = npq$ | $\sqrt{\frac{kpq}{n}}$ |
| Bernoulli | p | $p(1-p) = pq$ | \sqrt{pq} |

k = number of trials
n = overall sample size

General setup: place insects onto one of two diet formulations (A, B), measure if they molt (Yes/No)

Diet A



Diet B



Binomial trials (**5 insects per cup**)

Experiment option 1

Diet A



Diet B



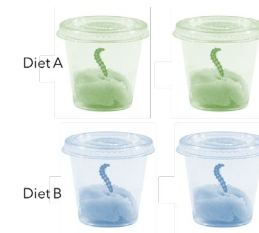
Bernoulli trials (**one insect per cup**)

Experiment option 2



| cup_id | n_molting | n_total | diet |
|--------|-----------|---------|------|
| 1 | 4 | 5 | A |
| 2 | 2 | 5 | A |
| 3 | 3 | 5 | B |
| 4 | 4 | 5 | B |

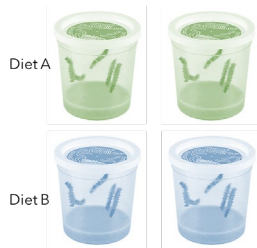
| cup_id | insect_id | diet | molted |
|--------|-----------|------|--------|
| 1 | 1 | A | Yes |
| 1 | 2 | A | Yes |
| 1 | 3 | A | Yes |
| 1 | 4 | A | Yes |
| 1 | 5 | A | No |
| 2 | 1 | A | Yes |
| 2 | 2 | A | No |
| 2 | 3 | A | No |
| 2 | 4 | A | No |
| 2 | 5 | A | Yes |
| | | | |



| cup_id | insect_id | molted | diet |
|--------|-----------|--------|------|
| 1 | 1 | Yes | A |
| 2 | 2 | Yes | A |
| 3 | 3 | No | B |
| 4 | 4 | Yes | B |

R code

```
df$molted_num <- ifelse(df$molted == "Yes", 1, 0)
fit1 <- glm(molted_num~diet, data=df,
family=binomial(link=logit))
```



| cup_id | n_molting | n_total | diet |
|--------|-----------|---------|------|
| 1 | 4 | 5 | A |
| 2 | 2 | 5 | A |
| 3 | 3 | 5 | B |
| 4 | 4 | 5 | B |

```
fit2 <- glm(cbind(n_molting, n_total-
n_molting)~diet, data=df,
family=binomial(logit))
```

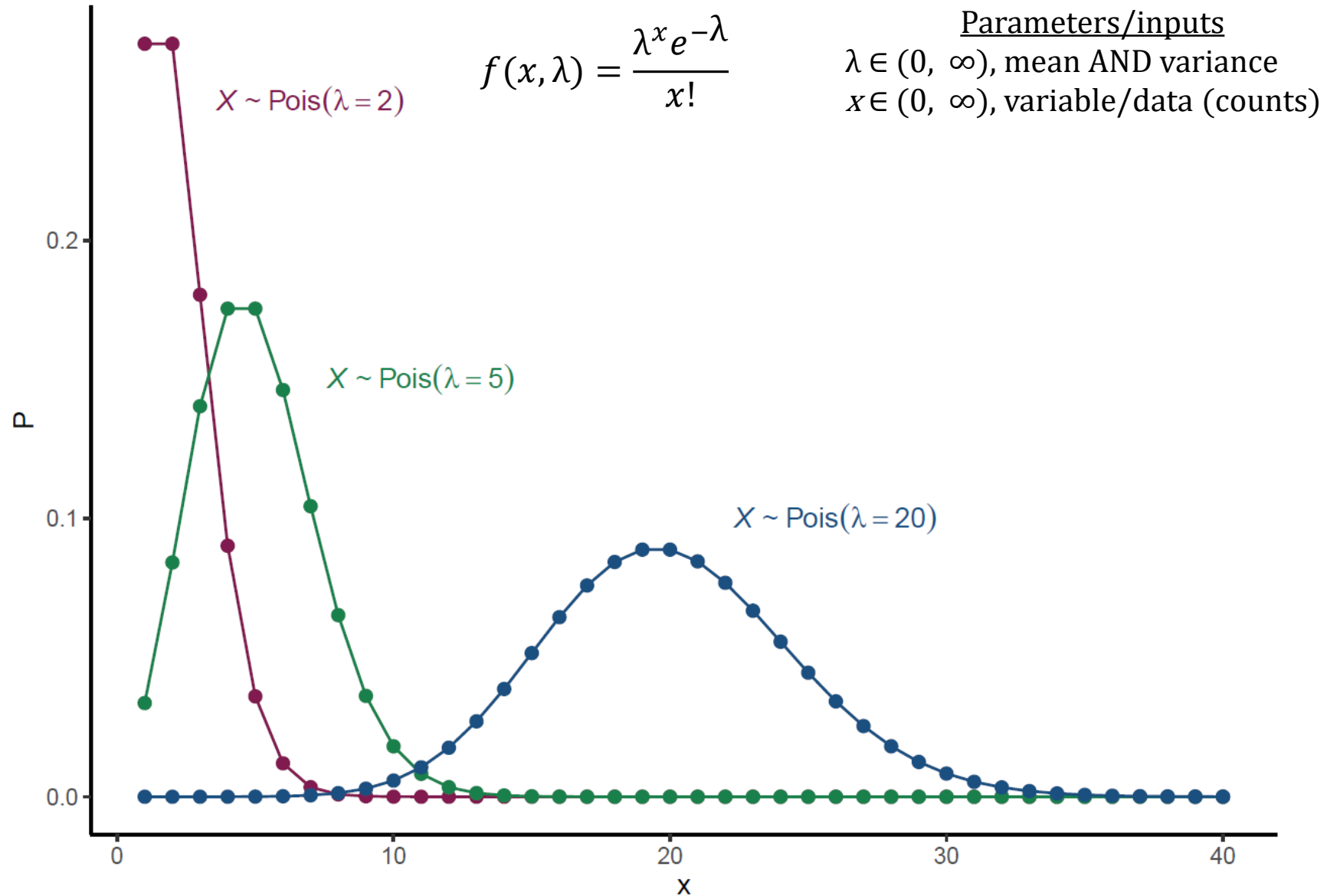
```
df$proportion_molting <- df$n_molting/df$n_total
fit2a <- glm(proportion_molting~diet, data=df,
family=binomial(logit), weights=n_total)
```

| cup_id | insect_id | diet | molted |
|--------|-----------|------|--------|
| 1 | 1 | A | Yes |
| 1 | 2 | A | Yes |
| 1 | 3 | A | Yes |
| 1 | 4 | A | Yes |
| 1 | 5 | A | No |
| 2 | 1 | A | Yes |
| 2 | 2 | A | No |
| 2 | 3 | A | No |
| 2 | 4 | A | No |
| 2 | 5 | A | Yes |
| | | | |

```
library(lme4)
df$molted_num <- ifelse(df$molted == "Yes", 1, 0)
fit3 <- glmer(molted_num~diet+(1|cup_id), data=df,
family=binomial(logit))
```

I am just showing you *inputs* for code...we will cover inputs, outputs, diagnostics, AND interpretations when we cover "generalized linear models" later in the semester

Poisson distribution



Poisson distribution



Poisson distribution



Placed mated, female moths individually onto one of two plant species, measured number of eggs laid.


| plant_id | n_eggs | plant_species |
|----------|--------|---------------|
| 1 | 10 | A |
| 2 | 22 | A |
| 3 | 23 | A |
| 4 | 4 | B |
| 5 | 205 | B |
| 6 | 59 | B |
| | | |

```
fit1 <- glm(n_eggs~plant_species, data=df,  
family=poisson(link=log))
```

Summary of distributions and their parameters

| Distribution | Mean | Variance | SE |
|--------------|-----------|-----------------|-----------------------------|
| Normal | μ | σ^2 | $\sqrt{\frac{\sigma^2}{n}}$ |
| Binomial | np | $np(1-p) = npq$ | $\sqrt{\frac{kpq}{n}}$ |
| Bernoulli | p | $p(1-p) = pq$ | \sqrt{pq} |
| Poisson | λ | λ | $\sqrt{\frac{\lambda}{n}}$ |

k = number of trials
n = overall sample size



NOTE: This is my assumption for what *should* be reported when presenting standard errors for Poisson distributed data, but (to my knowledge) people usually report the mean (λ ; the calculation is the same as for Gaussian/Normal data) but then they also use the SE formula as if the data were Gaussian; that is where the calculations would differ.

Distributions

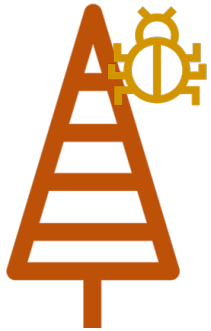
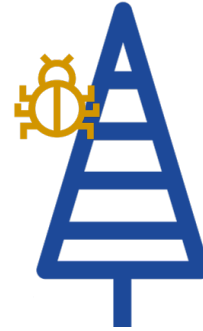
Identifying the distribution of a response variable is foundational to modeling, as it will determine the analysis we use and guide diagnostics.

Diagnostics: determining if the model is appropriate (e.g., does it provide a “sufficient” fit to the data?) and if the assumptions of the analysis are met.

We can't *know* the true distribution of a variable/population per se, so we make an informed guess/decision.

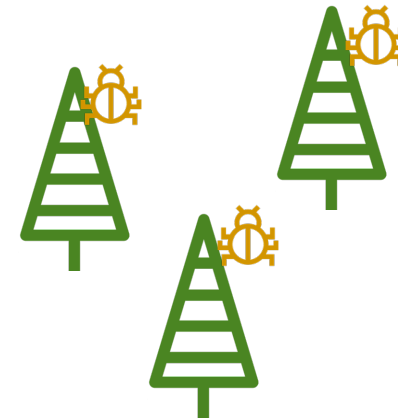
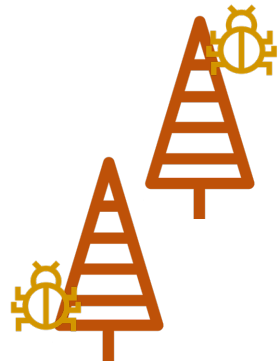
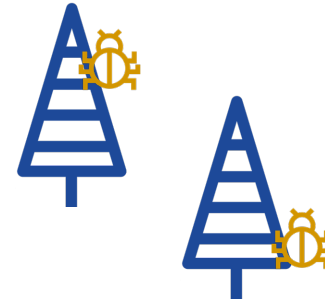
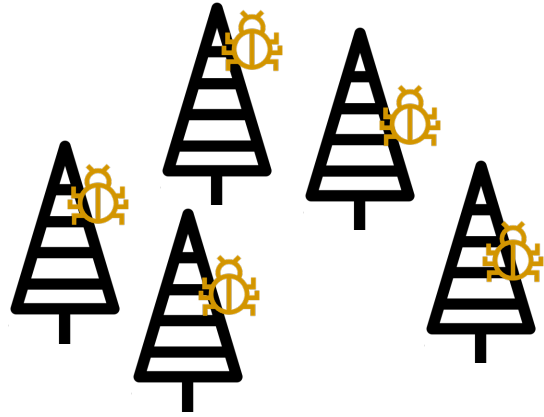
In a lot of linear models, we assume the residuals/errors are normally distributed...but we **also** often assume they are independent...

Independence



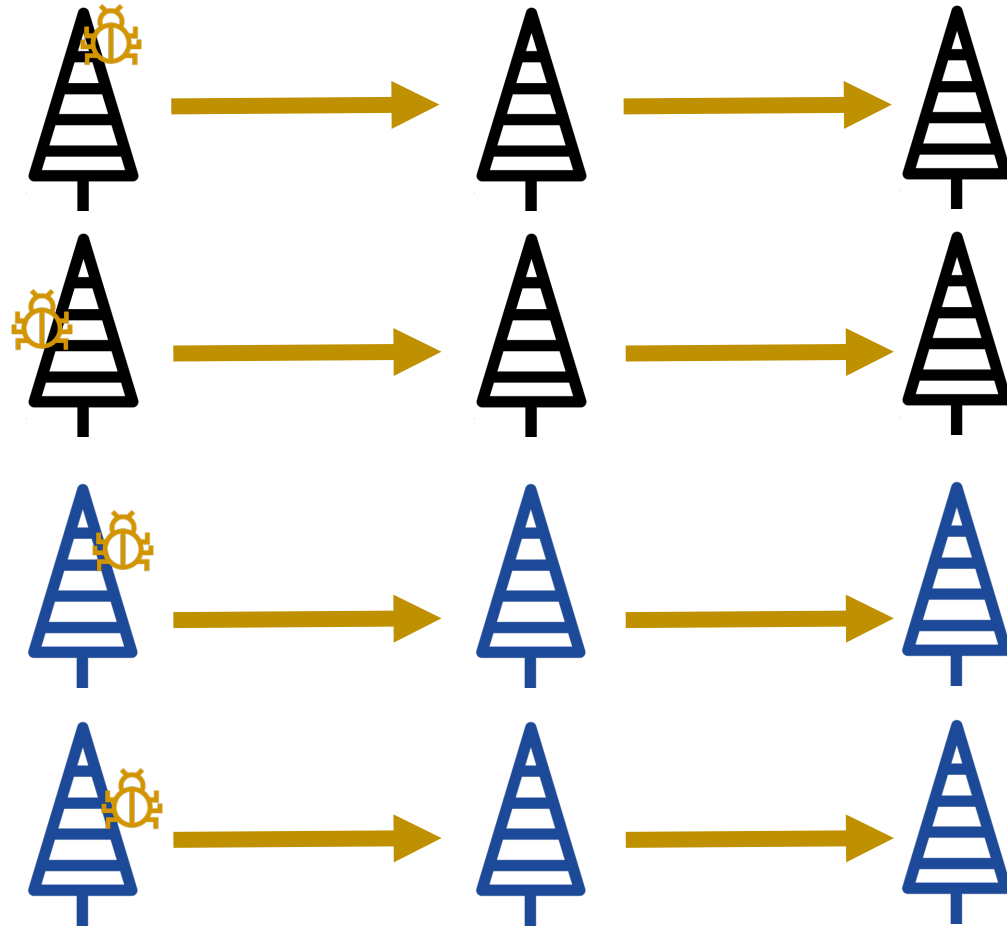
Each color represents a site. Each site has one insect sampled per tree.

Independence...?



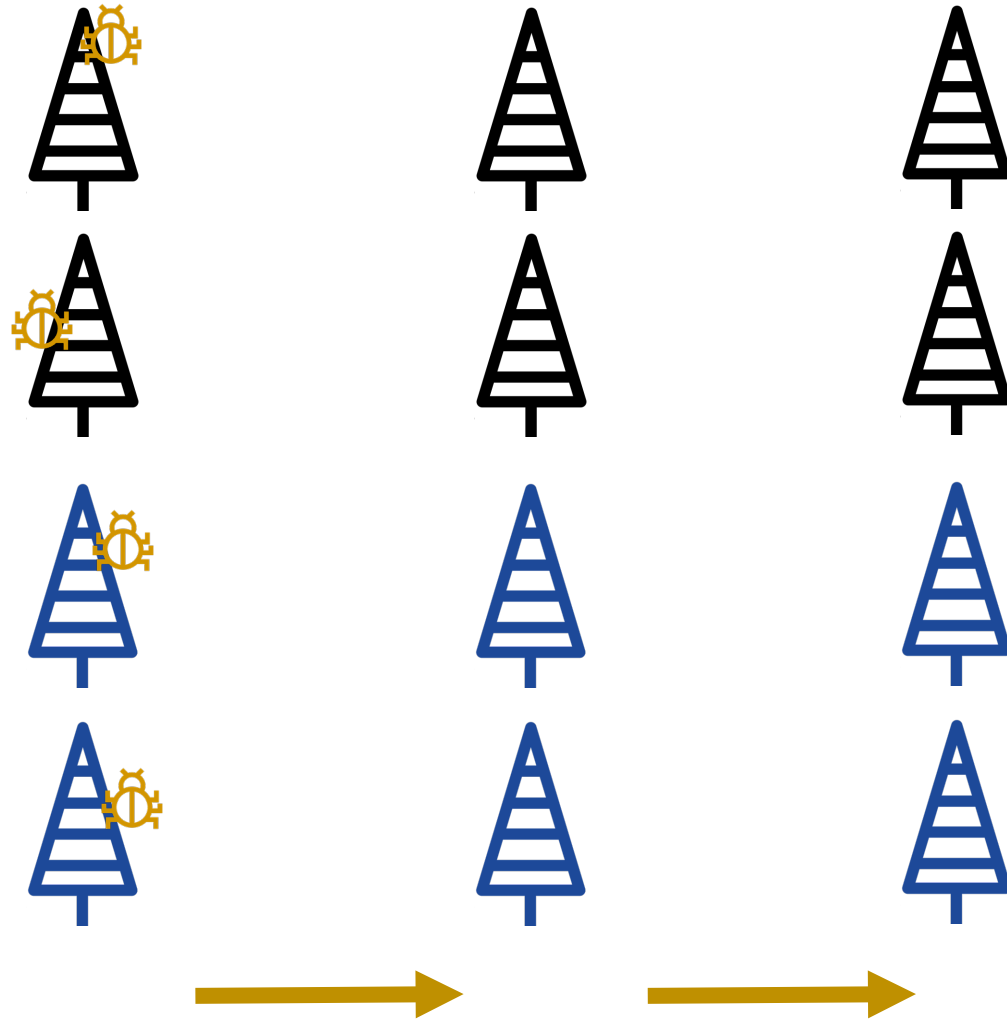
Each color represents a site. Each site has one insect sampled per tree, BUT multiple trees are sampled per site.

Repeated measures



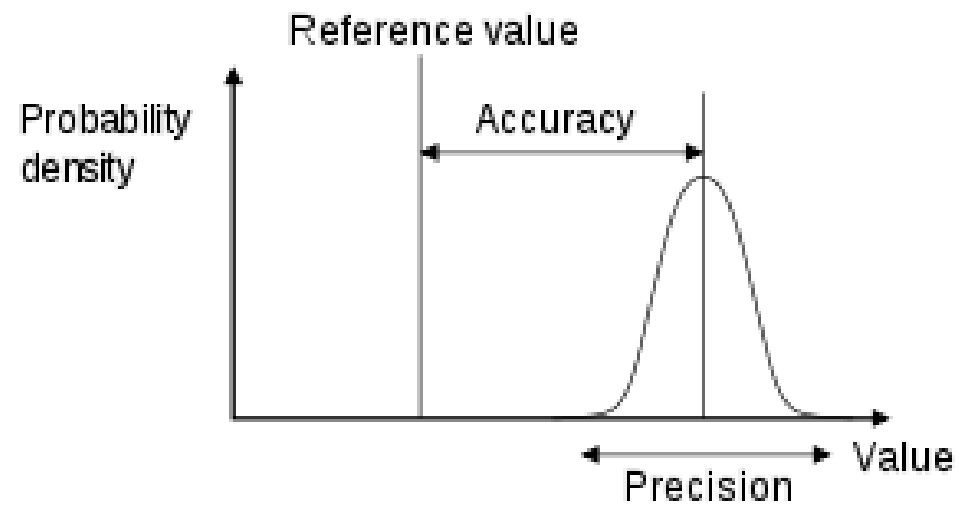
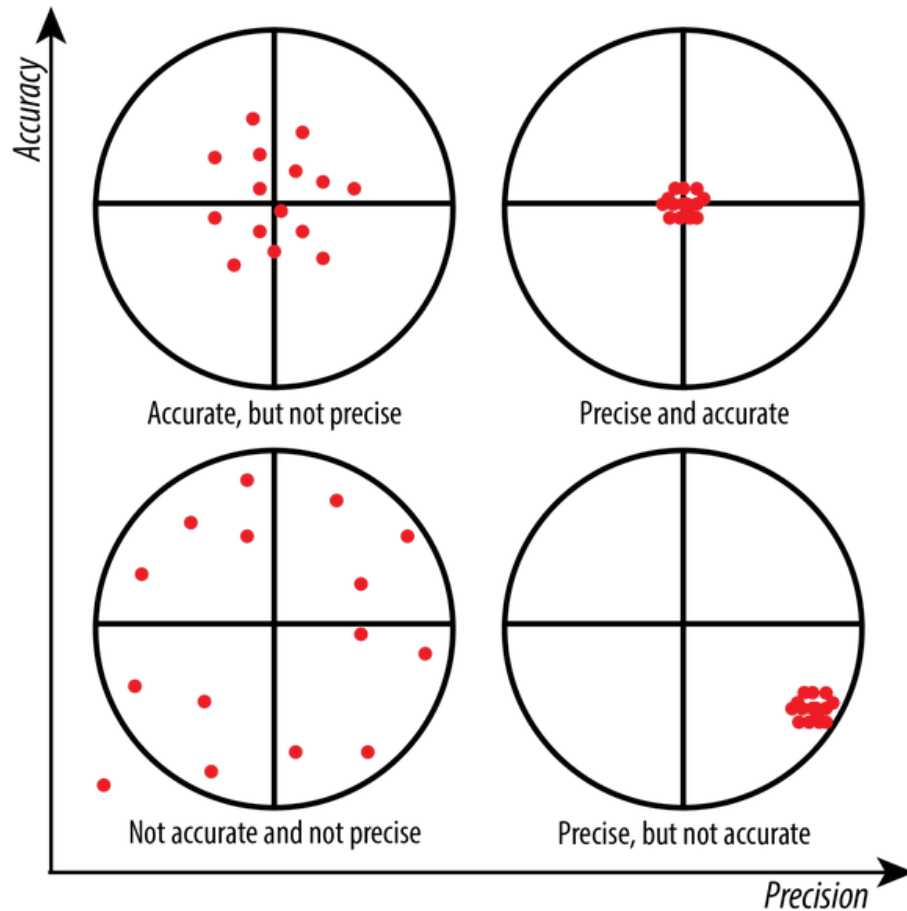
Each color represents a site. Each site has one sample per tree per day, BUT trees are sampled on multiple days.

Repeated measures



Each color represents a site. Each site has one sample per tree per day, BUT NEW trees are sampled on each visit.

Precision and accuracy



Four options in hypothesis testing

| | | Reality | |
|----------------|----------|---|---|
| | | Positive | Negative |
| Study findings | Positive | True positive (Power) $(1 - \beta)$ | False positive Type I error (α) |
| | Negative | False negative Type II error (β) | True negative |

Type I error

Type I error (α): rejecting the null hypothesis when it is actually true (false positive). You will frequently see in manuscripts something like "...we used a type I error rate of $\alpha=0.05$." Meaning that if a p -value is less than 0.05 in *your* study, you have "statistical significance" or "statistical clarity".

The correct interpretation of a p -value: it is the probability that you observed your data/results if the null hypothesis were true. So, a small p -value means that your results are really unlikely if the null hypothesis is true. In such cases, instead of concluding we just got unlikely results, we reject the null hypothesis.

Type II error

Type II error (β): failing to reject the null hypothesis when it is actually false (false negative).

The probability of committing a type II error is expressed as beta, β .

Power = $1 - \beta$ = probability of correctly rejecting the null hypothesis. A common target power = 0.80.

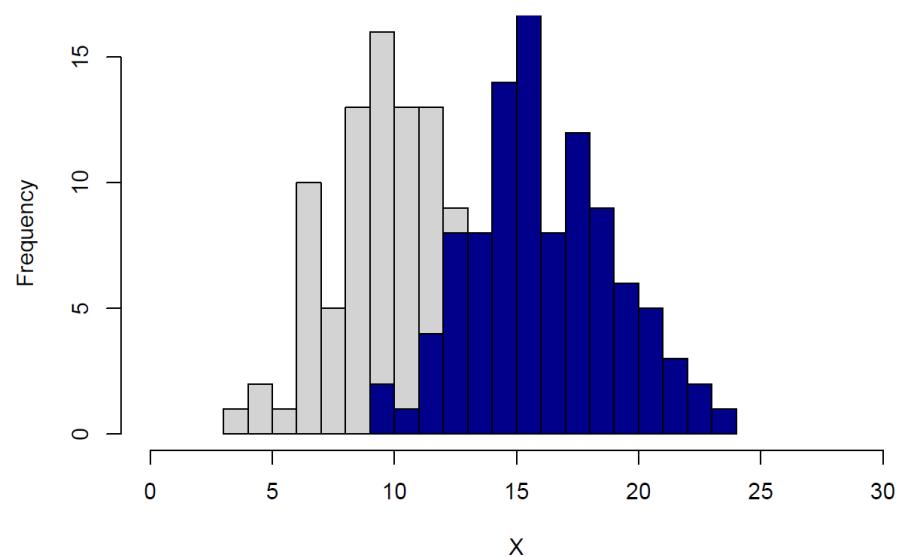
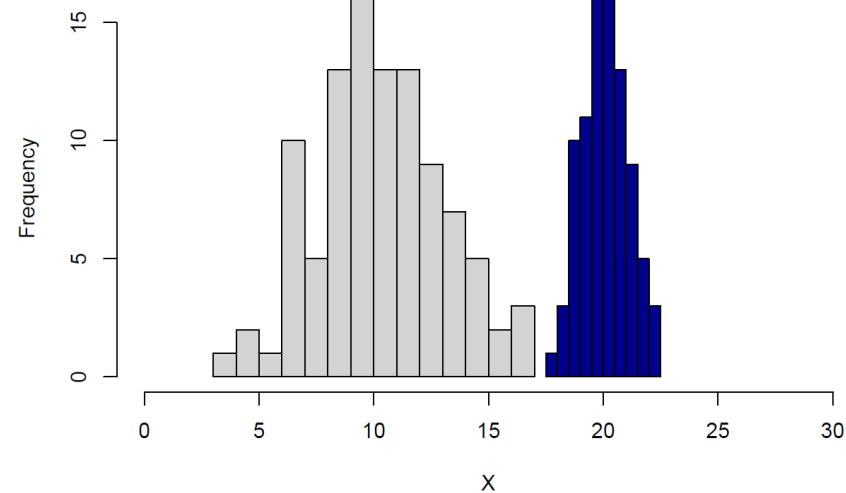
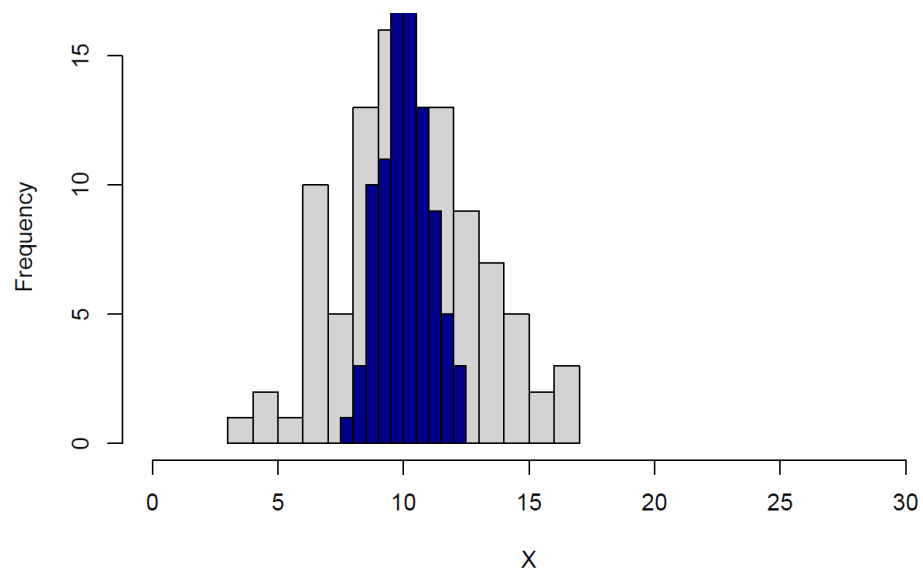
Power is a combination of sample size, sample variability, and the difference between sample means (or the strength of their correlation).

If two populations are different but highly variable, you would need “large” samples to differentiate them statistically.

An applet by Russ Lenth (U of Iowa) you can download and “prove” the concept:

<http://homepage.divms.uiowa.edu/~rlenth/Power/>

Power is largely determined by mean and variance



Activity: power of two sample t -test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Consider the t -value above. Remember: larger t -values \rightarrow smaller p -values \rightarrow and often happier scientists (generally...but please don't worship p -values).

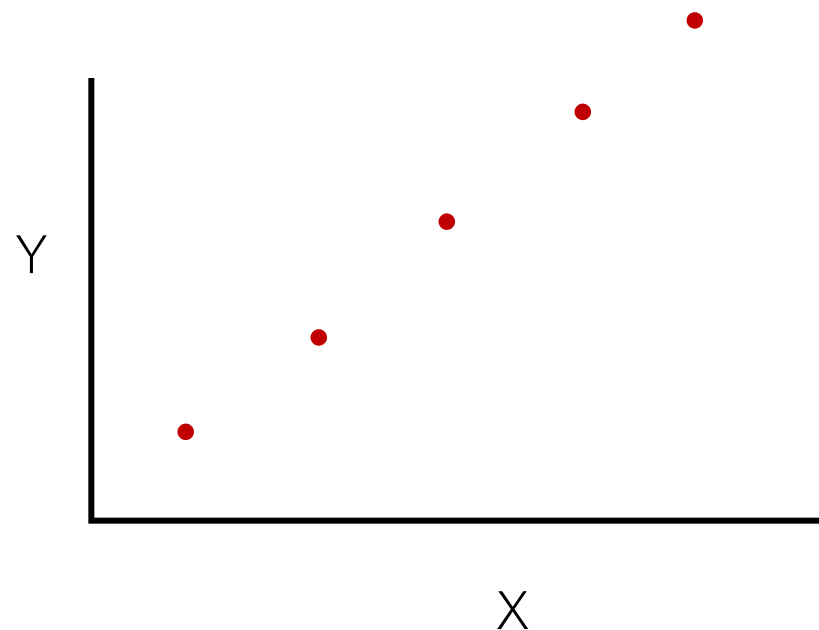
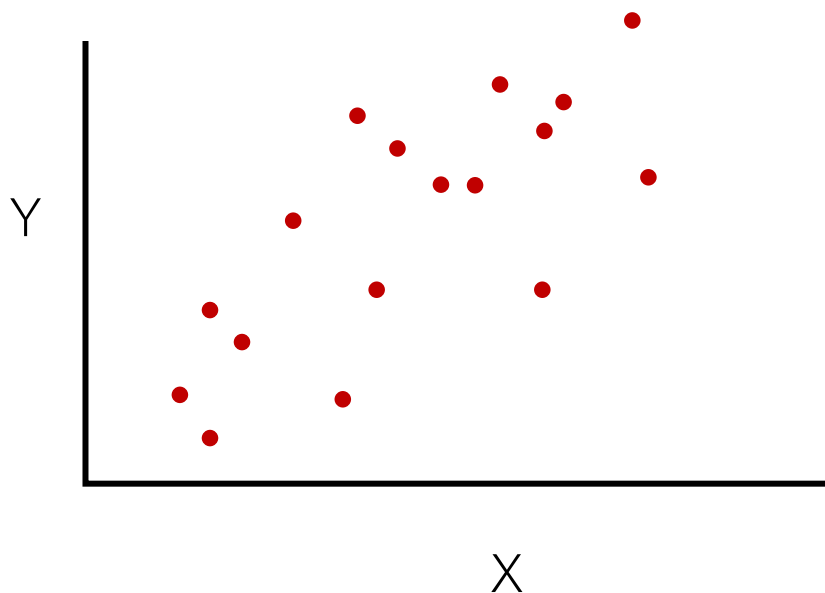
Think about the values below and write down what would happen to t as each approached infinity (i.e., got really big) AND all else was held equal:

$$\begin{array}{l} \bar{X}_1 - \bar{X}_2 \\ s_1^2 \\ N_2 \end{array}$$

Power in regression

The more variance, the larger sample size that is needed to detect an effect (if present)...but be careful not to “*p*-hack” (more on this in a second)

Think about the effect size and use your expertise to determine if it is meaningful.



Power analysis in R

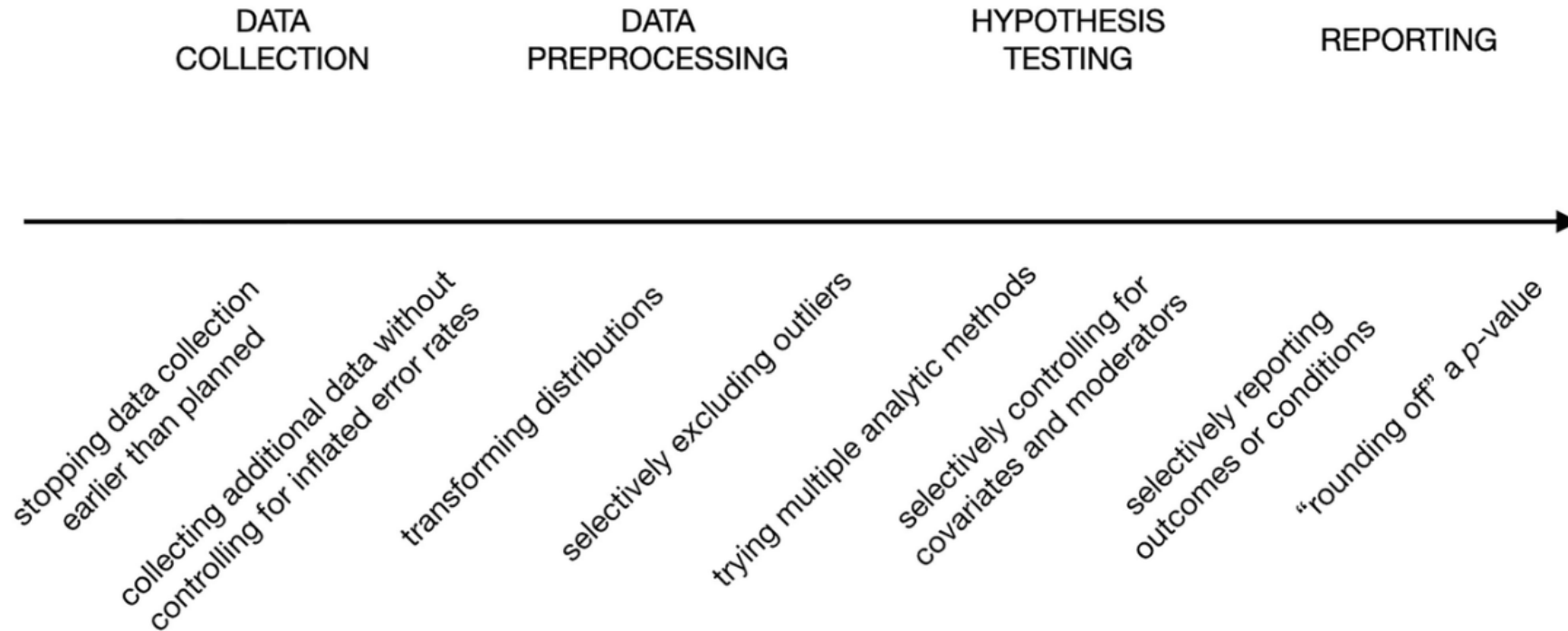
Functions in pwr

| Name ↕ | Description ↕ |
|----------------------------------|--|
| plot.power.htest | Plot diagram of sample size vs. test power |
| pwr.2p.test | Power calculation for two proportions (same sample sizes) |
| pwr.2p2n.test | Power calculation for two proportions (different sample sizes) |
| pwr-package | Basic Functions for Power Analysis pwr |
| pwr.t2n.test | Power calculations for two samples (different sizes) t-tests of means |
| ES.w1 | Effect size calculation in the chi-squared test for goodness of fit |
| pwr.t.test | Power calculations for t-tests of means (one sample, two samples and paired samples) |
| pwr.f2.test | Power calculations for the general linear model |
| ES.h | Effect size calculation for proportions |
| pwr.p.test | Power calculations for proportion tests (one sample) |
| ES.w2 | Effect size calculation in the chi-squared test for association |
| pwr.anova.test | Power calculations for balanced one-way analysis of variance tests |
| pwr.chisq.test | power calculations for chi-squared tests |

p -hacking

Fig. 5.1

From: The Myriad Forms of p -Hacking



Schematic depiction of exemplary p -hacking practices according to when they typically occur during the research process

Reis, D., Friese, M. (2022). The Myriad Forms of p -Hacking. In: O'Donohue, W., Masuda, A., Lilienfeld, S. (eds) Avoiding Questionable Research Practices in Applied Psychology. Springer, Cham.
https://doi.org/10.1007/978-3-031-04968-2_5