https://statisticsbyjim.com/basics/probability-distributions/

Comprendre les distributions de probabilités

Par Jim Frost 39 Commentaires

Une distribution de probabilité est une fonction qui décrit la probabilité d'obtenir les valeurs possibles qu'une variable aléatoire peut supposer. En d'autres termes, les valeurs de la variable varient en fonction de la distribution de probabilité sous-jacente.

Supposons que vous tiriez un échantillon aléatoire et mesuriez les hauteurs des sujets. Lorsque vous mesurez des hauteurs, vous pouvez créer une distribution des hauteurs. Ce type de distribution est utile lorsque vous devez savoir quels résultats sont les plus probables, la répartition des valeurs potentielles et la probabilité de résultats différents.

Dans cet article de blog, vous découvrirez les distributions de probabilité pour les variables discrètes et continues . Je vais vous montrer comment ils fonctionnent et des exemples d'utilisation.

Propriétés générales des distributions de probabilités

Les distributions de probabilité indiquent la probabilité d'un événement ou d'un résultat. Les statisticiens utilisent la notation suivante pour décrire les probabilités:

p (x) = la probabilité qu'une variable aléatoire prenne une valeur spécifique de x.

La somme de toutes les probabilités pour toutes les valeurs possibles doit être égale à 1. De plus, la probabilité pour une valeur particulière ou une plage de valeurs doit être comprise entre 0 et 1.

Les distributions de probabilité décrivent la dispersion des valeurs d'une variable aléatoire. Par conséquent, le type de variable détermine le type de distribution de probabilité. Pour une seule variable aléatoire, les statisticiens divisent les distributions selon les deux types suivants:

- Distributions de probabilité discrètes pour les variables discrètes
- Fonctions de densité de probabilité pour les variables continues

Vous pouvez utiliser des équations et des tableaux de valeurs de variables et de probabilités pour représenter une distribution de probabilités. Cependant, je préfère les représenter graphiquement à l'aide de diagrammes de distribution de probabilité. Comme vous le verrez dans les exemples qui suivent, les différences

entre les distributions de probabilité discrète et continue sont immédiatement apparentes. Vous verrez pourquoi j'aime ces graphiques!

Article connexe : Types de données et comment les utiliser

Distributions de probabilités discrètes

Les fonctions de probabilité discrètes sont également appelées fonctions de masse de probabilité et peuvent prendre un nombre discret de valeurs. Par exemple, les lancers de pièces et le nombre d'événements sont des fonctions discrètes. Ce sont des distributions discrètes car il n'y a pas de valeurs intermédiaires. Par exemple, vous ne pouvez avoir que des têtes ou des queues dans un tirage au sort. De même, si vous comptez le nombre de livres qu'une bibliothèque extrait par heure, vous pouvez compter 21 ou 22 livres, mais rien entre les deux.

Pour les fonctions de distribution de probabilité discrètes, chaque valeur possible a une probabilité non nulle. De plus, les probabilités pour toutes les valeurs possibles doivent être égales à un. La probabilité totale étant de 1, l'une des valeurs doit se produire pour chaque opportunité.

Par exemple, la probabilité de lancer un nombre spécifique sur un dé est de 1/6. La probabilité totale pour les six valeurs est égale à un. Lorsque vous lancez un dé, vous obtenez inévitablement l'une des valeurs possibles.

Si la distribution discrète a un nombre fini de valeurs, vous pouvez afficher toutes les valeurs avec leurs probabilités correspondantes dans un tableau. Par exemple, selon une étude, la probabilité du nombre de voitures dans un ménage californien est la suivante:

Number of Cars	Probability
0	0.03
1	0.13
2	0.70
3	0.10
4+	0.04

Types de distribution discrète

Il existe une variété de distributions de probabilité discrètes que vous pouvez utiliser pour modéliser différents types de données. La distribution discrète correcte dépend des propriétés de vos données. Par exemple, utilisez:

- Distribution binomiale pour modéliser des données binaires, telles que des lancers de pièces.

- Distribution de Poisson pour modéliser les données de comptage, telles que le nombre de sorties de livres de bibliothèque par heure.
- Distribution uniforme pour modéliser plusieurs événements avec la même probabilité, comme lancer un dé.

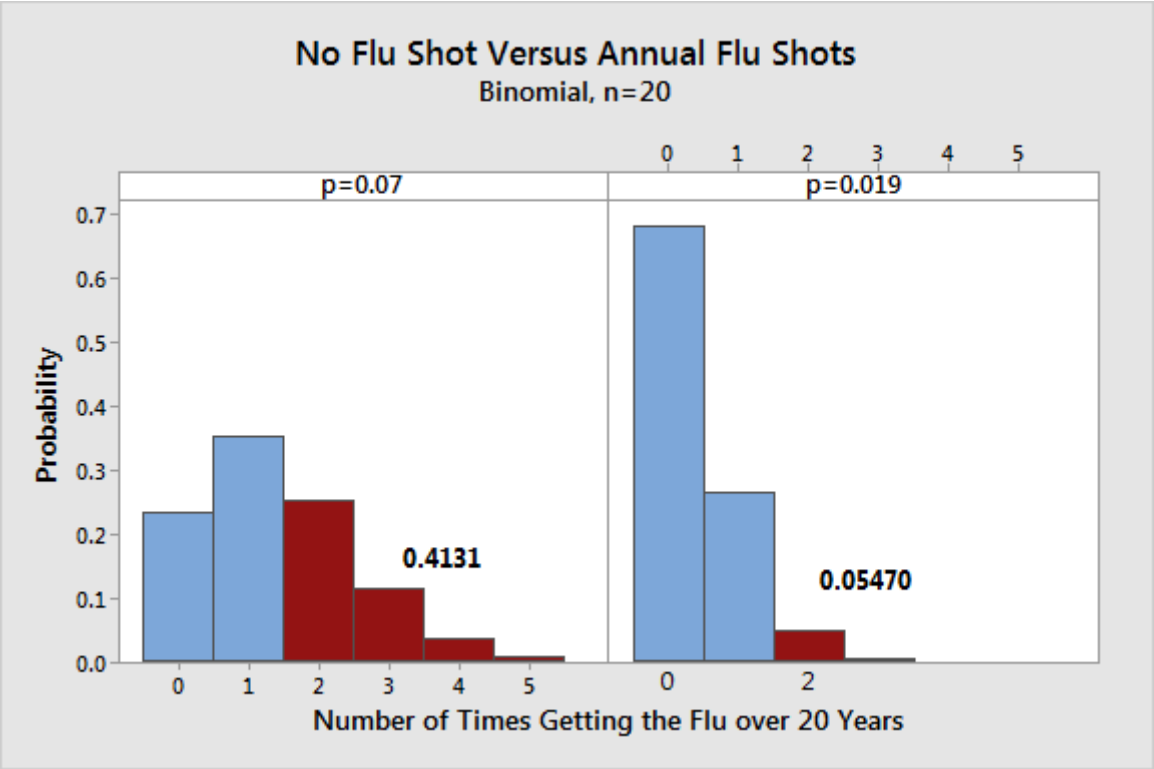
Pour en savoir plus sur plusieurs distributions de probabilités que vous pouvez utiliser avec des données binaires, lisez mon article Maximiser la valeur de vos données binaires .

Pour savoir comment déterminer si une distribution discrète spécifique est appropriée pour vos données, lisez mon post Tests d'adéquation des distributions discrètes .

Exemple d'utilisation de distributions de probabilités discrètes

Tous les exemples que j'inclus dans ce post vous montreront pourquoi j'aime représenter graphiquement les distributions de probabilité. Le cas ci-dessous provient de mon article de blog qui présente une analyse statistique de l'efficacité du vaccin antigrippal . J'utilise la distribution binomiale pour répondre à la question: combien de fois puis-je m'attendre à attraper la grippe pendant 20 ans avec ou sans vaccinations annuelles?

Cet exemple utilise des données binaires car les deux résultats possibles sont soit infectés par la grippe, soit non infectés par la grippe. Sur la base de diverses études, la probabilité à long terme d'une infection grippale est de 0,07 par an pour les non vaccinés et de 0,019 pour les vaccinés. Le graphique branche ces probabilités dans la distribution binomiale pour afficher le modèle de résultats pour les deux scénarios sur vingt ans. Chaque barre indique la probabilité d'attraper la grippe le nombre de fois spécifié. De plus, j'ai ombré les barres en rouge pour représenter la probabilité cumulée d'au moins deux infections grippales en 20 ans. Le panneau de gauche affiche les résultats attendus sans vaccination tandis que le panneau de droite montre les résultats avec les vaccinations annuelles.



Une différence significative vous saute aux yeux, ce qui démontre la puissance des diagrammes de distribution de probabilité! La barre la plus grande du graphique est celle du panneau de droite qui représente zéro cas de grippe en 20 ans lorsque vous recevez des vaccins contre la grippe. Lorsque vous vaccinez chaque année, vous avez 68% de chances de ne pas attraper la grippe dans les 20 ans! À l'inverse, si vous ne vaccinez pas, vous n'avez que 23% d'échappant entièrement à la grippe.

Dans le panneau de gauche, la distribution s'étale beaucoup plus loin que dans le panneau de droite. Sans vaccination, vous avez 41% de chances de contracter la grippe au moins deux fois en 20 ans, contre 5% avec les vaccinations annuelles. Certains malchanceux non vaccinés attraperont la grippe quatre ou cinq fois au cours de cette période!

Distributions de probabilités continues

Les fonctions de probabilité continues sont également appelées fonctions de densité de probabilité. Vous savez que vous avez une distribution continue si la variable peut prendre un nombre infini de valeurs entre deux valeurs quelconques. Les variables continues sont souvent des mesures sur une échelle, comme la taille, le poids et la température.

Contrairement aux distributions de probabilité discrètes où chaque valeur particulière a une probabilité non nulle, les valeurs spécifiques dans les distributions continues ont une probabilité nulle. Par exemple, la probabilité de mesurer une température qui est exactement de 32 degrés est nulle.

Pourquoi? Considérez que la température peut être un nombre infini d'autres températures infinitésimalement supérieures ou inférieures à 32. Les statisticiens disent qu'une valeur individuelle a une probabilité infinitésimalement petite équivalente à zéro.

Comment trouver des probabilités pour des données continues

Les probabilités de distributions continues sont mesurées sur des plages de valeurs plutôt que sur des points uniques. Une probabilité indique la probabilité qu'une valeur tombe dans un intervalle. Cette propriété est simple à démontrer en utilisant un diagramme de distribution de probabilité - que nous verrons bientôt!

Sur un graphique de probabilité, l'aire entière sous la courbe de distribution est égale à 1. Ce fait équivaut à la façon dont la somme de toutes les probabilités doit être égale à une pour les distributions discrètes. La proportion de l'aire sous la courbe qui se situe dans une plage de valeurs le long de l'axe X représente la probabilité qu'une valeur se situe dans cette plage. Enfin, vous ne pouvez pas avoir une zone sous la courbe avec une seule valeur, ce qui explique pourquoi la probabilité est égale à zéro pour une valeur individuelle.

Caractéristiques des distributions de probabilités continues

Tout comme il existe différents types de distributions discrètes pour différents types de données discrètes, il existe différentes distributions pour les données continues . Chaque distribution de probabilité a des paramètres qui définissent sa forme. La plupart des distributions ont entre 1 et 3 paramètres. La spécification de ces paramètres établit entièrement la forme de la distribution et toutes ses probabilités. Ces paramètres représentent des propriétés essentielles de la distribution, telles que la tendance centrale et la variabilité.

Related posts : Comprendre les mesures de tendance centrale et comprendre les mesures de variabilité

La distribution continue la plus connue est la distribution normale, également connue sous le nom de distribution gaussienne ou «courbe en cloche». Cette distribution symétrique s'adapte à une grande variété de phénomènes, tels que la taille humaine et les scores de QI. Il a deux paramètres: la moyenne et l'écart type. La distribution de Weibull et la distribution lognormale sont d'autres distributions continues courantes. Ces deux distributions peuvent correspondre à des données asymétriques .

Les paramètres de distribution sont des valeurs qui s'appliquent à des populations entières. Malheureusement, les paramètres de population sont généralement

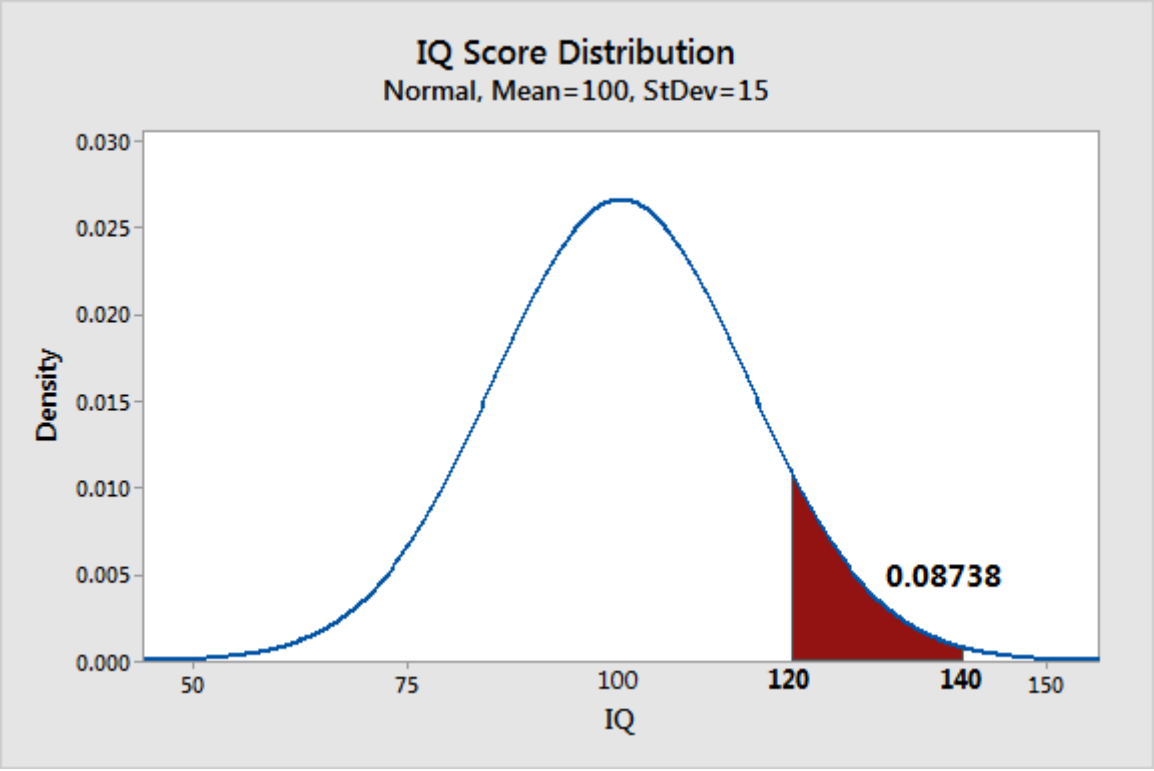
inconnus car il est généralement impossible de mesurer une population entière. Cependant, vous pouvez utiliser des échantillons aléatoires pour calculer les estimations de ces paramètres.

Pour savoir comment déterminer la distribution qui correspond le mieux à vos exemples de données, lisez mon article sur Comment identifier la distribution de vos données .

Exemple d'utilisation de la distribution de probabilité normale

Commençons par la distribution normale pour montrer comment utiliser les distributions de probabilité continues.

La distribution des scores de QI est définie comme une distribution normale avec une moyenne de 100 et un écart-type de 15. Nous allons créer le graphique de probabilité de cette distribution. De plus, déterminons la probabilité qu'un score de QI se situe entre 120 et 140.



Examinez les propriétés du graphique de probabilité ci-dessus. Nous pouvons voir que c'est une distribution symétrique où les valeurs se produisent le plus souvent autour de 100, ce qui est la moyenne. Les probabilités diminuent lorsque vous vous éloignez de la moyenne dans les deux directions. La zone ombrée pour la plage de scores de QI entre 120 et 140 contient 8,738% de la surface totale sous la

courbe. Par conséquent, la probabilité qu'un score de QI se situe dans cette plage est de 0,08738.

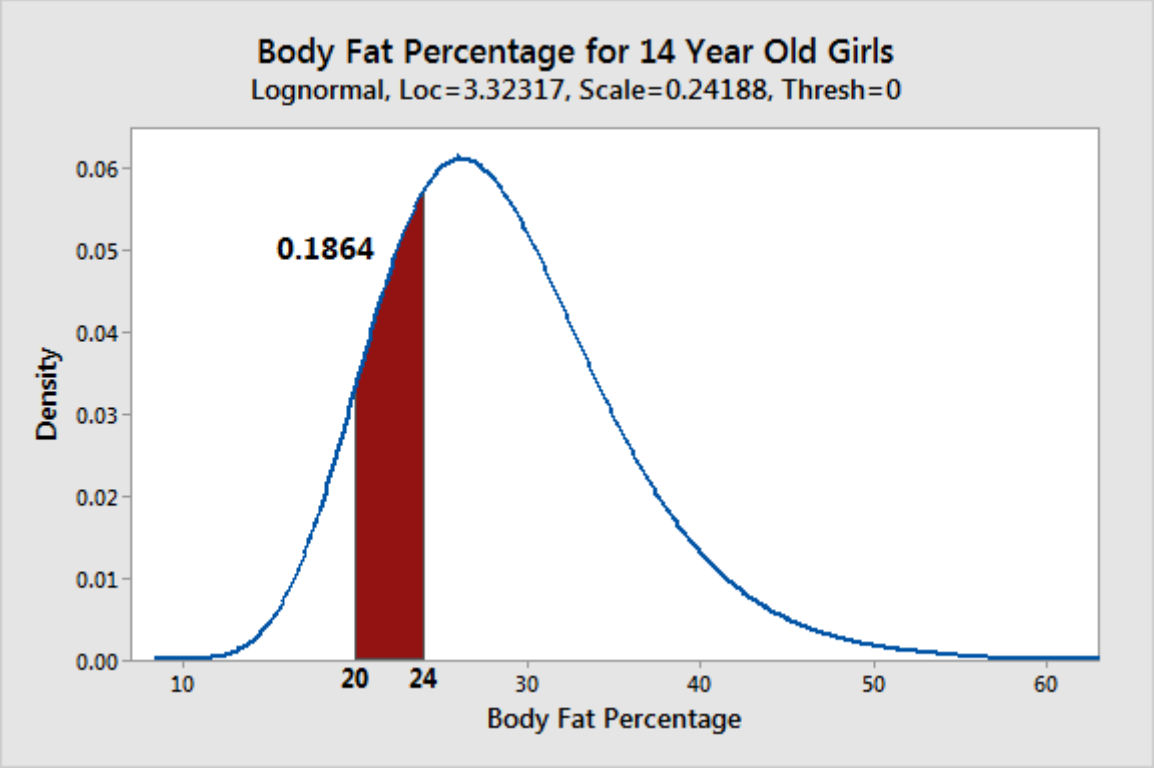
Article connexe : Utilisation de la distribution normale

Exemple d'utilisation de la distribution de probabilité lognormale

Comme je l'ai mentionné, j'aime beaucoup les diagrammes de distribution de probabilité car ils rendent les propriétés de distribution claires. Dans l'exemple ci-dessus, nous avons utilisé la distribution normale. Parce que cette distribution est si bien connue, vous avez peut-être deviné l'apparence générale du graphique. Voyons maintenant un exemple moins intuitif.

Supposons qu'on vous dise que les pourcentages de graisse corporelle pour les adolescentes suivent une distribution lognormale avec un emplacement de 3,32317 et une échelle de 0,24188. De plus, vous êtes invité à déterminer la probabilité que les valeurs de pourcentage de graisse corporelle se situent entre 20 et 24%. Hein? La forme de cette distribution n'est probablement pas claire, les valeurs les plus courantes et la fréquence à laquelle les valeurs se situent dans cette plage!

La plupart des logiciels statistiques vous permettent de tracer des distributions de probabilités et de répondre à toutes ces questions à la fois.



Le graphique affiche à la fois la forme de la distribution et la façon dont notre gamme d'intérêt s'y intègre. Nous pouvons voir qu'il s'agit d'une distribution asymétrique et les valeurs les plus courantes tombent près de 26%. De plus, notre plage d'intérêt tombe en dessous du pic de la courbe et contient 18,64% des occurrences.

Comme vous pouvez le voir, ces graphiques sont un moyen efficace de signaler des informations de distribution complexes à un public non averti.

Cette distribution fournit le meilleur ajustement pour les données que j'ai recueillies pour une étude. Découvrez comment j'ai identifié la distribution de ces données .

Le test d'hypothèse utilise des distributions de probabilité spéciales

Le test d'hypothèse statistique utilise des types particuliers de distributions de probabilité pour déterminer si les résultats sont statistiquement significatifs. Plus précisément, ils utilisent des distributions d'échantillonnage et les distributions de statistiques de test .

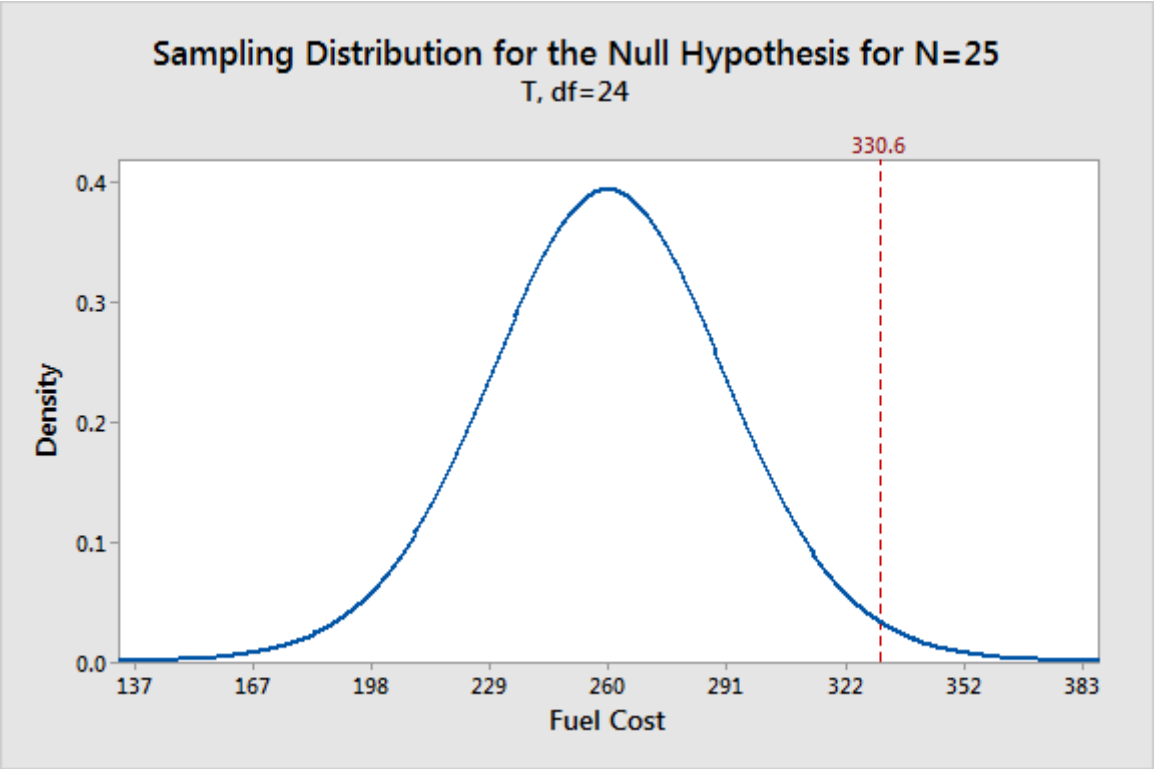
Distribution d'échantillonnage

Un concept essentiel dans les statistiques inférentielles est que l'échantillon aléatoire particulier que vous tirez pour une étude n'est qu'un parmi un grand nombre d'échantillons possibles que vous auriez pu tirer de votre population d'intérêt. La compréhension de ce contexte plus large de tous les échantillons possibles et de la manière dont l'échantillon de votre étude s'y intègre fournit des informations précieuses.

Supposons que nous prélevions un nombre substantiel d'échantillons aléatoires de même taille dans la même population et calculons la moyenne de l'échantillon pour chaque échantillon. Au cours de ce processus, nous observons un large éventail de moyennes d'échantillons et nous pouvons représenter graphiquement leur distribution.

Ce type de distribution est appelé distribution d'échantillonnage. Les distributions d'échantillonnage vous permettent de déterminer la probabilité d'obtenir différentes valeurs d'échantillon, ce qui les rend cruciales pour effectuer des tests d'hypothèse .

Le graphique ci-dessous montre la distribution d'échantillonnage des coûts énergétiques. Il montre quelles moyennes d'échantillon sont plus et moins susceptibles de se produire lorsque la moyenne de la population est de 260. Il affiche également la moyenne d'échantillon spécifique qu'une étude obtient (330,6). Le graphique indique que la moyenne de notre échantillon observé n'est pas la valeur la plus probable, mais elle n'est pas non plus totalement invraisemblable. Les tests d'hypothèse utilisent ce type d'informations pour déterminer si les résultats sont statistiquement significatifs.

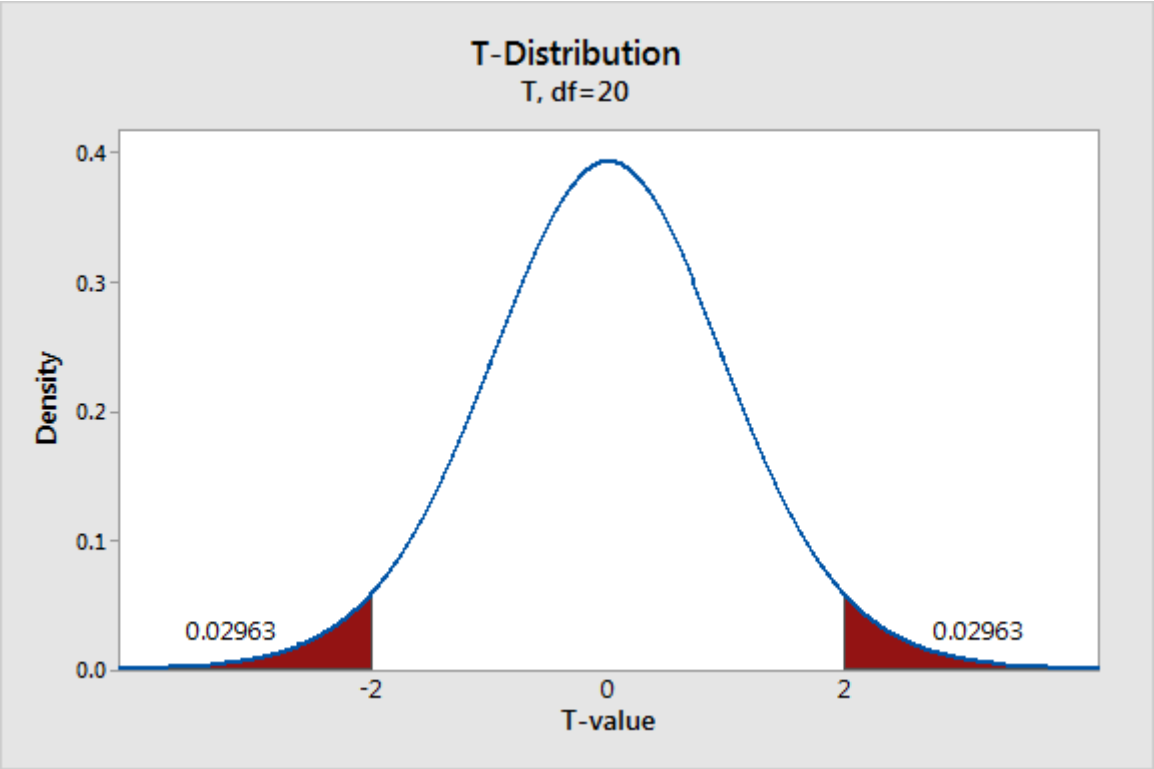


Pour en savoir plus sur l'échantillonnage des distributions, lisez mon article sur le fonctionnement des tests d'hypothèse .

Distributions pour les statistiques de test

Chaque type de test d'hypothèse utilise une statistique de test. Par exemple, les tests t utilisent des valeurs t, l'ANOVA utilise des valeurs F et les tests chi carré utilisent des valeurs chi carré. Les tests d'hypothèse utilisent les distributions de probabilité de ces statistiques de test pour calculer les valeurs de p. C'est vrai, les valeurs p proviennent de ces distributions!

Par exemple, un test t prend toutes les données de l'échantillon et les résume à une seule valeur t , puis la distribution t calcule la valeur p . Le graphique de distribution de probabilité ci-dessous représente un test t bilatéral qui produit une valeur t de 2. Le graphique de la distribution t indique que chacune des deux régions ombrées qui correspond à des valeurs t de +2 et -2 (c'est l'aspect bilatéral du test) a une probabilité de 0,02963 - pour un total de 0,05926. C'est la valeur de p pour ce test!



Pour en savoir plus sur la façon dont cela fonctionne pour différents tests d'hypothèse, lisez mes articles sur:

- Comment fonctionnent les tests t
- Fonctionnement du test F dans l'ANOVA unidirectionnelle
- Degrés de liberté (il y a une section sur les distributions de probabilité.)

J'espère que vous pouvez voir à quel point les distributions de probabilités sont cruciales dans les statistiques et pourquoi je pense que les représenter graphiquement est un moyen puissant de transmettre les résultats!

Si vous apprenez les statistiques et aimez l'approche que j'utilise dans mon blog, consultez mon livre électronique Introduction aux statistiques !